MKM – ein Metamodell für Korpusmetadaten

Dokumentation und Wiederverwendung historischer Korpora

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

doctor philosophiae

(Dr. phil.)

eingereicht an
der Sprach- und literaturwissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M.A. Carolin Odebrecht

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst
Dekanin der Sprach- und literaturwissenschaftlichen Fakultät
Prof. Dr. Ulrike Vedder

Gutachterin/Gutachter:

1. Prof. Dr. Anke Lüdeling, Humboldt-Universität zu Berlin
2. Dr. Laurent Romary, INRIA, Frankreich

Datum der Verteidigung: 21.07.2017

All Models are wrong but some are useful

(Box 1979: 202)

Abstracts

Deutsche Fassung

Korpusdokumentation wird in dieser Arbeit als eine Voraussetzung für die Wiederverwendung von Korpora und als ein Bestandteil des Forschungsdatenmanagements verstanden, welches unter anderem die Veröffentlichung und Archivierung von Korpora umfasst. Verschiedene Forschungsdaten stellen ganz unterschiedliche Anforderungen an die Dokumentation und können auch unterschiedlich wiederverwendet werden. Ein geeignetes Anwendungsbeispiel stellen historische Textkorpora dar, da sie in vielen Fächern als empirische Grundlage für die Forschung genutzt werden können. Sie zeichnen sich im Weiteren durch vielfältige Unterschiede in ihrer Aufbereitung und durch ein komplexes Verhältnis zu der historischen Vorlage aus. Die Ergebnisse von Transkription und Normalisierung müssen als eigenständige Repräsentationen und Interpretationen im Vergleich zur Vorlage verstanden werden. Was müssen Forscherinnen und Forscher über ihr Korpus mit Hilfe von Metadaten dokumentieren, um dessen Erschließung und Wiederverwendung für andere Forscherinnen und Forscher zu ermöglichen? Welche Funktionen übernehmen dabei die Metadaten? Wie können Metadaten modelliert werden, um auf alle Arten von historischen Korpora angewendet werden zu können? Die Arbeit und ihre Fragestellung sind fest in einem interdisziplinären Kontext verortet. Für die Beantwortung der Forschungsfragen wurden Erkenntnisse und Methoden aus den Fachbereichen der Korpuslinguistik, der historischen Linguistik, der Informationswissenschaft sowie der Informatik theoretisch und empirisch betrachtet und für die Entwicklung eines Metamodells für Korpusmetadaten fruchtbar gemacht. Das im Rahmen dieser Arbeit in UML entwickelte Metamodell für Korpusmetadaten modelliert Metadaten von historischen textbasierten Korpora aus einer technisch-abstrakten, produktorientierten und überfachlichen Perspektive und ist in einer TEI-Spezifikation mit Hilfe der TEI-eigenen Modellierungssprache ODD realisiert.

English Version

Corpus documentation is a requirement for enabling corpus reuse scenarios and is a part of research data management which covers, among others, data publication and archiving. Different types of research data make differing demands on corpus documentation, and may be reused in various ways. Historical corpora represent an interesting and challenging use case because they are the foundation for empirical studies in many disciplines and show a great variety of reuse possibilities, of data creation, and of data annotation. Furthermore, the relation between the historical corpus and the historical original is complex. The transcription and normalisation of historical texts must be understood as independent representations and interpretations in their own right. Which kind of metadata information, then, must be included in a corpus documentation in order to enable intellectual access and reuse scenarios? What kind of role do metadata play? How can metadata be designed to be applicable to all types of historical corpora? These research questions can only be addressed with help of an interdisciplinary approach, considering findings and methods of corpus linguistics, historical linguistics, information science and computer science. The metamodel developed in this thesis models metadata of historical text-based corpora from a technical, abstract, and interdisciplinary point of view with help of UML. It is realised as a TEI-specification using the modelling language ODD.

Danksagung

Anke Lüdeling und Laurent Romary möchte ich sehr herzlich für die hervorragende Betreuung, ihre Unterstützung und die wertvollen Fachgespräche danken. Anke hat mir korpuslinguistische Methoden zur Datenerstellung und -analyse beigebracht und mich bei der anspruchsvollen Arbeit mit Korpora von Nichtstandardvarietäten begleitet. Durch Laurent habe ich die Arbeit mit den Frameworks der TEI und dabei insbesondere deren Modellebenen kennengelernt sowie mich mit unterschiedlichen Perspektiven des Forschungsdatenmanagements und unterschiedlichen Textkonzeptionen auseinandergesetzt. Die Arbeit ist im Rahmen des DFG-geförderten Projektes LAUDATIO entstanden, aber auch immer Teil der Arbeitsgruppe Korpuslinguistik und Morphologie der Humboldt-Universität zu Berlin gewesen.
Ich möchte mich ebenfalls bei meinen Kolleginnen und Kollegen für die vielfältigen Anregungen, Fachgespräche sowie ihre Kommentare und Anmerkungen zu einzelnen Kapiteln meiner Arbeit bedanken. Ihre jeweiligen fachlichen Perspektiven haben mich sehr unterstützt. Vielen lieben Dank an Euch (Reihenfolge randomisiert): Amir Zeldes, Gohar Schnelle, Thomas Krause, Florian Zipser, Stephan Druskat, Laura Perlitz, Hagen Hirschmann und Vivian Voigt. Des Weiteren möchte ich folgenden Personen für einen wertvollen Austausch zu Themen dieser Arbeit danken (Reihenfolge randomisiert): Cerstin Mahlow, Kerstin Eckart, Michael Piotrowski und Svetlana Petrova.
Weiterhin möchte ich mich bei meinem Bruder Thomas Odebrecht für seine wertvollen Kommentare zu Kapiteln meiner Arbeit bedanken. Die uneingeschränkte Unterstützung meiner Familie war für mich in jeder Phase dieser Arbeit wichtig und dafür möchte ich mich bei Euch allen bedanken. Schließlich möchte ich Malte Belz danken, der mir durch seine fachliche und emotionale Unterstützung besonderen Rückhalt gegeben hat.

```
<?xml version="1.0" encoding="UTF-8"?>
<danksagung an="euchAlle">
    Vielen lieben Dank!
</danksagung>
```

Inhaltsverzeichnis

1 Zielstellung und Forschungsfrage

Die vorliegende Arbeit befasst sich mit den Voraussetzungen der Wiederverwendung von historischen Korpora und stellt dabei folgende Forschungsfrage: Wie können historische Korpora für ein überfachliches Publikum dokumentiert werden, so dass diese ersteller- und fachunabhängig zum Zweck der Wiederverwendung erschlossen werden können?

Für die Beantwortung dieser Forschungsfrage wird untersucht, welche Eigenschaften Korpora besitzen, und welche davon in einer Korpusdokumentation für andere Forscherinnen und Forscher[1] beschrieben werden müssen. Metadaten übernehmen dann die Aufgabe der Dokumentation. Daraus ergeben sich weitere Forschungsfragen: Wie können Metadaten eine Menge von historischen Korpora beschreiben? Welche Informationen über Korpora sind für die Wiederverwendung relevant? Was können Forscherinnen und Forscher über ihr Korpus dokumentieren, um dessen Erschließung und Wiederverwendung für andere Forscherinnen und Forscher ermöglichen zu können?

Diese Forschungsfragen sind in den Forschungskontext der Dokumentation, Veröffentlichung und Archivierung von Forschungsdaten eingebettet, in dem die Anforderungen für die Nachhaltigkeit von Forschungsdaten (und von Software) identifiziert und von Initiativen wie Data Archiving and Networked Services (DANS) mit dem Data Seal of Approval[2] oder den FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016) erforscht und in allgemeine Richtlinien formuliert werden. In diesem Rahmen stellen sich auch viele Projekte wie Digital Research Infrastructure for the Arts and Humanities (DARIAH)[3] (Romary und Chambers 2014) und Common Language Resources and Technology Infrastructure (CLARIN) [4] (Hinrichs

[1] In dieser Arbeit richte ich mich in den Fließtextformulierungen nach der Vorgabe zur geschlechtergerechten Sprache der Humboldt-Universität zu Berlin, vgl. `https://www.hu-berlin.de/de/service/online/websites/richtlinien/styleguide/geschlechtergerechte-sprache`. Aus Gründen der Übersichtlichkeit wird die Binnen-I-Schreibung in den Abbildungen verwendet.

[2] `http://www.datasealofapproval.org` (besucht am 27.01.2017).

[3] `http://www.dariah.eu` (besucht am 23.01.2017).

[4] `https://www.clarin.eu` (besucht am 23.01.2017).

und Krauwer 2014) sowie Initiativen und Fachgemeinschaften wie das Linguistic Data Consortium (LDC)[5], die Text Encoding Initiative (TEI)[6], die Dublin Core Metadata Initiative (DMCI)[7], die Open Language Archives Community (OLAC)[8] oder die Research Data Alliance (RDA)[9] den gleichen Herausforderungen für die unterschiedlichsten Forschungsdatentypen und -anwendungen.[10]

Diese Arbeit setzt sich speziell mit den Anforderungen einer Korpusdokumentation auseinander, die als eine Voraussetzung für die Wiederverwendung von Korpora und als ein Bestandteil der Veröffentlichung und Archivierung von Korpora verstanden werden kann. Weiterhin fokussiert die Arbeit auf den speziellen Forschungsdatentyp `historisches Textkorpus`. Historische Textkorpora eignen sich als Untersuchungsgegenstand besonders gut, da sie in vielen Fächern als empirische Grundlage der Forschung genutzt werden. Darüber hinaus zeichnen sich die historischen Korpora durch starke Unterschiede in ihrer Realisierung von historischen Texten aus. Das Verhältnis zwischen historischer Vorlage und Digitalisat ist hoch komplex und wird ganz unterschiedlich umgesetzt. So kann ein historisches Korpus auch als digitale Edition, als Textsammlung ober Belegsammlung interpretiert werden. Dies hat auch einen großen Einfluss auf die Korpusarchitektur, die z. T. wesentlich komplexer als bei modernen Textkorpora gestaltet ist. Ein solches komplexes Geflecht aus verschiedenen Konzepten für Text und verschiedenen Korpusarchitekturen jeweils zu dokumentieren, stellt eine besondere Herausforderung bei der Erstellung einer Korpusdokumentation dar. Da historische Korpora in dieser Hinsicht einen besonders komplizierten Fall darstellen, eigenen sie sich besonders als Gegenstand dieser Arbeit.

Nun soll nicht nur ein Vorschlag einer Korpusdokumentation für ein einzelnes konkretes historisches Korpus erarbeitet werden, sondern ein Vorschlag, der auf den Korpustyp `historisches Textkorpus` allgemein anwendbar ist. Damit muss über die Eigenschaften vorhandener historischer Korpora abstrahiert werden, um jeweils diese Eigenschaften in einem Modell abbilden zu können. Das Ziel ist es, über verschiedene einzelne Eigenschaftsmodelle historischer Korpora zu abstrahieren und eine gemeinsame einheitliche Beschreibungsebene darüber zu modellieren. Eine solche Abstraktion von Eigenschaften historischer Korpora, wie sie hier notwendig ist, wird den

[5] `https://www.ldc.upenn.edu/` (besucht am 23.01.2017).
[6] `http://www.tei-c.org/` (besucht am 23.01.2017).
[7] `http://http://dublincore.org/` (besucht am 23.01.2017).
[8] `http://www.language-archives.org` (besucht am 23.01.2017).
[9] `https://www.rd-alliance.org` (besucht am 23.01.2017).
[10] Eine ausführliche Kontextualisierung der Arbeit und Diskussion bisheriger Ansätze erfolgt in Kapitel 5.

bisherigen Ansätzen zur Dokumentation von Forschungsdaten nicht oder nur teilweise zugrunde gelegt. Ein weiteres Ziel ist, bislang nicht bekannte Korpora ebenfalls damit beschreiben zu können. Mit einem solchen Metamodell für Korpusmetadaten wird dann die Voraussetzung für eine Wiederverwendung von historischen Korpora durch eine einheitliche, erstellerunabhängige und extensive Dokumentation geschaffen werden. Das Metamodell leistet darüber hinaus einen theoretisch-methodischen Beitrag zum Forschungsdatum `historisches Textkorpus` in den Bereichen der korpusbasierten Forschung und insbesondere der Korpuslinguistik.

1.1 Erschließung von historischen Korpora

Korpora stellen allgemein in vielen Geisteswissenschaften wie der Linguistik, der Geschichtswissenschaft oder der Literaturwissenschaft die empirische Grundlage der Forschung dar.[11] Die korpusbasierte Forschung hat ihre Anfänge bereits in den 1940er Jahren und wird als Methode vielfältig und fächerübergreifend weiterentwickelt (Lüdeling und Zeldes 2007). Ein gemeinsamer, überfachlicher Ausgangspunkt ist dafür die zugrundeliegende sprachliche Ressource. Im Fall der historischen Korpora sind das historische Texte.

So werden beispielsweise in der Linguistik historische Zeitungstexte als Korpora auf verschiedene Arten aufbereitet und als empirische Grundlage für verschiedenen Untersuchungen genutzt, wie das Mannheimer Korpus Historischer Zeitungen und Zeitschriften (IDS 2013), die Mercurius-Baumbank (Demske 2005, 2007) oder das German Manchester Corpus (GerManC) (Bennett et al. 2007; Durrell et al. 2007). Alle beispielhaft genannten Korpora beinhalten Texte aus Zeitungen des 16.–19. Jahrhundert, die in den jeweiligen Projekten unterschiedlich digitalisiert und in verschiedenen Formaten mit unterschiedlichen Kategorisierungen für beispielsweise *Wortart* oder *Satz* annotiert sind.

Nicht nur innerhalb eines Fachs werden gleiche oder vergleichbare sprachliche Ressourcen erstellt: Historische Privatbriefe werden beispielsweise als empirische Grundlage genutzt und als `Korpus`[12] aufbereitet, wie z. B. in der Pädagogik die Gesamtausgabe der Briefe Fröbels[13], in der Literaturwissenschaft die Briefe und Texte aus dem intellektuellen Berlin um 1800 (Baillot und Seifert 2013)

[11] Vgl. für die Methoden der Digitalisierung in den textbasierten Wissenschaften z. B. Haugen und Apollon (2014).
[12] Zur Definition des Begriffs `Korpus` und die Unterscheidung zu Editionen vgl. Kapitel 2.
[13] `http://bbf.dipf.de/digitale-bbf/editionen/froebel/ausgabe` (besucht am 21.12.2016).

oder in der historischen Linguistik das FÜRSTINNENKORRESPONDENZKORPUS (Lühr et al. 2014).

Wenn nun Forscherinnen und Forscher eines dieser Korpora wiederverwenden wollen, müssen sie sich dieses Korpus erschließen, sich also mit dem Korpus, den enthaltenen Texten und der Art der Aufbereitung vertraut machen. Das jeweilige Korpus muss dann aus Sicht der Forscherinnen und Forscher entweder innerhalb eines Fachs oder überfachlich erschlossen werden.

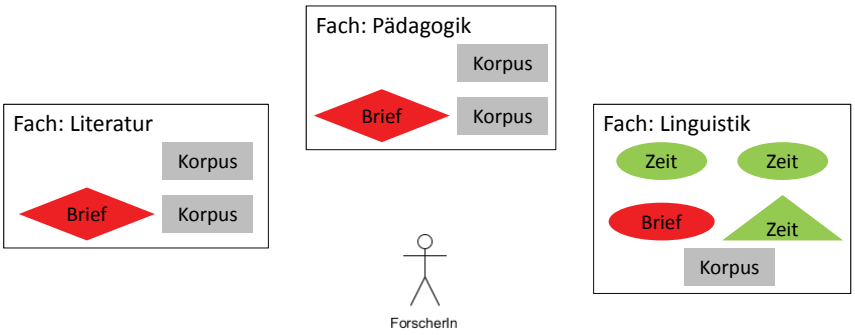

Abbildung 1.1: *Inner- und überfachliche Erschließung von Korpora durch Forscherinnen und Forscher mit einem jeweils eigenen fachlichen Zugang. Die im Korpus enthaltene Textsorte ist angegeben. Die verschiedenen Formen illustrieren, dass die Korpora verschieden aufbereitet sind.*

Abbildung 1.1 illustriert die Erschließung von Korpora in Abhängigkeit vom Fach, in dem die Korpora erstellt werden. Korpora aus Zeitungstexten (hier grün „Zeit“), Korpora aus Briefen (hier rot „Brief“) sowie alle denkbaren weiteren Korpora aus den unterschiedlichsten sprachlichen Ressourcen (hier grau „Korpus“) werden den verschiedenen Fachgebieten zugeordnet. Da die Korpora unterschiedlich aufbereitet, also z. B. verschieden annotiert sind und damit auch unterschiedliche Eigenschaften besitzen, werden sie mit verschiedenen Formen dargestellt (Ellipse, Dreieck, Rechteck und Raute). Forscherinnen und Forscher stehen dann vor der Herausforderung, das jeweilige Korpusangebot pro Fach zu erschließen oder erst nach Korpora zu durchsuchen. Bei der Suche nach Korpora sind weitere wichtige Kriterien *recall*, *precision* und *access*:

> For instance, the user may not be able to find all the existing data about the language of interest because different sites have called it by different

> names (low recall). The user may be swamped with irrelevant resources because search terms have important meanings in other domains (low precision). The user may not be able to use an accessible data file for lack of being able to match it with the right tools. (Bird und Simons 2001: 8)

Ein geringer *Rücklauf* („recall“) meint, dass Forscherinnen und Forscher nicht alle relevanten Korpora finden können. Andersherum meint eine geringe *Präzision* („precision“), dass zu viele, nicht relevante Korpora gefunden werden. Der Zugang („access“) zu den gefunden Korpora ist auch nicht immer frei oder erschwert, so dass die Korpora nicht gut erschlossen werden können.

Forscherinnen und Forscher, die beispielsweise auf Grundlage des FÜRSTINNENKORRESPONDENZKORPUS (Linguistik, Brief) linguistische Phänomene untersucht haben und nun ihre empirische Grundlage erweitern wollen, müssen sich dann Korpora anderer Fächer wie der Literaturwissenschaft erschließen, um die Geeignetheit der Korpora zur Wiederverwendung im Rahmen ihrer eigenen Forschung zu prüfen. Dieselben historischen Zeitungstexte, die beispielsweise als linguistische Korpora digital aufbereitet wurden, können auch literaturwissenschaftlich untersucht werden. Umgekehrt können ebenso kritische digitale Editionen der Literaturwissenschaft als Grundlagen für linguistische Sprachwandeluntersuchungen dienen. Beispielsweise kann auch eine historische Quelle in Form eines Korpus aus der Geschichtsgeographie als empirische Grundlage für die Untersuchung von historischen Sprachständen genutzt werden (Greenstein und Burnard 1995: 139). Diese Beispiele motivieren, dass Forscherinnen und Forscher in verschiedenen Fächern ihre Forschungsfragen sowohl an die gleichen als auch an unterschiedliche Textsorten stellen können. Um die empirische Grundlagen der eigenen Forschung mit Hilfe facheigener und fachfremder Korpora zu erweitern oder auf andere Sprecher oder Sprachgebiete auszuweiten, müssen also geeignete Korpora gefunden und erschlossen werden.

Diese unterschiedlichen Korpora können ebenfalls über einen gemeinsamen Zugang und nicht nur in einem jeweiligen fachbezogenen Kontext zur Verfügung stehen (Abbildung 1.2).

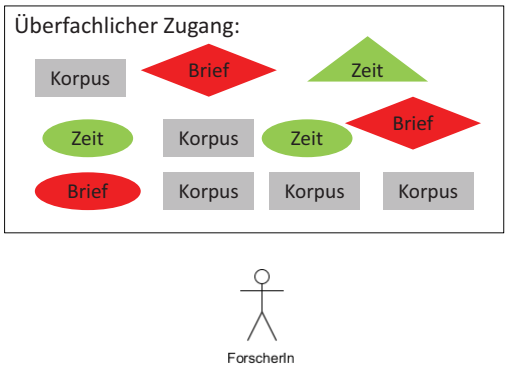

Abbildung 1.2: *Die Erschließung von Korpora durch Forscherinnen und Forscher mit einem überfachlichen Zugang. Die im Korpus enthaltene Textsorte ist angegeben. Die verschiedenen Formen illustrieren, dass die Korpora verschiedenen aufbereitet sind.*

Die Herausforderung in allen Fällen der Erschließung ist für die Forscherinnen und Forscher, die Strukturen und Inhalte fremder Korpora zu verstehen. Dabei kann die inner- wie überfachliche Erschließung von Korpora im Prinzip auf zwei Wegen erfolgen:

Erschließung über die Ressource: Die Korpora können in ihrem jeweiligen Format mit den dazu passenden Analyse- oder Annotationstools geöffnet, ausgelesen oder ausgewertet werden. Für ein Korpus in einem XML-basierten Format muss das Format erkannt und beherrscht werden und eine entsprechende Software zum Auslesen und Visualisieren der jeweiligen Formate installiert und bedient werden. Dadurch, dass es eine Vielzahl an verschiedenen Korpusarchitekturen und Annoationsarten in einem Fach und zwischen den Fächern gibt (vgl. Kapitel 2), erscheint eine format- und softwareabhängige Erschließung von Korpora sehr aufwändig. Somit käme unter Umständen pro Korpus nicht nur die Erschließung des Korpus selbst sondern auch die Erschließung von zusätzlich mindestens einem Format oder einer Formatspezifikation sowie einem Tool samt seiner Bedienung hinzu. Erkannt werden muss dann, wie die Annotationen und damit die Korpusarchitektur strukturell aufgebaut werden, welche Annotationskonzepte darin wie abgebildet sind und wie ein Tool dies den Nutzerinnen und Nutzern präsentiert.

Erschließung über die Korpusdokumentation: Viele Korpora besitzen eine Art von Dokumentation. Solche Dokumentationen liegen typischerweise in einer Art

Fließtext in Form von Homepages, Annotationsrichtlinien und Handbüchern sowie wissenschaftlichen Artikeln vor, die jeweils von den Forscherinnen und Forschern einzeln gelesen und ausgewertet werden müssen. Häufig sind solche Korpusdokumentationen nur indirekt untereinander vergleichbar, weil sie stark in ihrer Struktur und Aussagekraft variieren können und auf die einzelnen Forschungsgegenstände, Forschungsfragen oder deren Ergebnisse fokussiert sind. Die enthaltenen Informationen können damit nicht jedem Zweck vollständig oder ausreichend dienen. Diese Dokumentationen müssen damit eher einzeln, fach- und forschungsorientiert erschlossen und nach den gewünschten Informationen gefiltert werden. Ein Vorteil dieser Art der Erschließung ist, dass sich solche Informationen auf einer Metaebene zum Korpus befinden, also nicht deren integraler Bestandteil sind und damit nicht über Annotations- oder Analysetools ausgelesen werden müssen. Wenn die Dokumentationen nicht unabhängig vorliegen, dann müssen sie wiederum durch die Ressource (das Korpus) und ihre Realisierungen (Formate) interpretiert werden.

Solche Informationen über Korpora werden allgemein auch als `Metadaten` verstanden.[14] Der Begriff der `Metadaten` ist für die Korpusdokumentation also zentral. Metadaten können strukturierte oder unstrukturierte Informationen über ein Datum (hier Korpus) geben und sowohl separat vom Korpus als auch im Korpusformat selbst vorliegen. Mit der Erschließung von Korpora über eine Korpusdokumentation werden hier von den einzelnen Korpora unabhängige Dokumentationen durch Metadaten verstanden, die nicht direkt mit Korpusformaten oder -tools zusammen abgebildet werden (können).

Gegenstand dieser Arbeit ist daher die Erschließung von historischen Korpora über deren Metadaten.[15] Es wird in dieser Arbeit herausgearbeitet, wie und in welcher Form Metadaten helfen können, Korpora überfachlich zum Zweck der Wiederverwendung zu erschließen.

Eine Voraussetzung dafür, dass Forschungsdaten wiederverwendet werden können, ist natürlich, dass sie auch öffentlich zugänglich und frei zur Verfügung gestellt werden:

> In jedem Fall sollten die erhobenen Daten nach Abschluss der Forschun-

[14] Für eine genaue Definition und Einordnung von Metadaten vgl. Kapitel 4.

[15] Eine wissenschaftliche Auseinandersetzung mit den verschiedenen Annotationmodellen für Korpora oder einzelnen Annotationsformaten oder Analysetools für Korpora wird in dieser Arbeit nicht angestrebt.

> gen öffentlich zugänglich und frei verfügbar sein. Dieses ist die wesentliche Voraussetzung dafür, dass Daten im Rahmen neuer Fragestellungen wieder genutzt werden können sowie dafür, dass im Falle von Zweifeln an der Publikation die Daten für die Überprüfung der publizierten Ergebnisse herangezogen werden können. (Deutsche Forschungsgemeinschaft 2009: 2)

Wenn dies gelingt, dann kann beispielsweise der Aufwand einer – möglicherweise erneuten oder weiteren – Digitalisierung von derselben oder einer vergleichbaren Ressource vermieden werden oder die Notwendigkeit eines solchen Schritts kann mit einer Erschließung über eine Korpusdokumentation vorab geprüft werden. Im Bedarfsfall können bereits vorhandene Datenstrukturen nachgenutzt und für die eigene Forschung erweitert oder neu zusammengestellt werden.

Der Umgang mit Forschungsdaten in einem umfassenden Rahmen wird allgemein als `Forschungsdatenmanagement` verstanden, welches die Erstellung, die Auswertung, die Publikation und schließlich die Nachnutzung von Forschungsdaten regelt.[16]

Ein Ziel dieser Arbeit ist es, diese Anforderungen in Bezug auf die Dokumentation von Korpora miteinzubeziehen:

> Das Forschungsdatenmanagement muss so gestaltet werden, dass Datenzugriff und -auswertung unabhängig vom Datenerzeuger möglich wird und bleibt. Neben der technischen Speicherung und Lesbarkeit der Forschungsdaten müssen ausreichend Informationen zu ihrer Interpretation in Metadaten überliefert werden. (Büttner et al. 2011: 14)

Die Dokumentation von Korpora ist also Teil des Forschungsdatenmanagements. Eine Anreicherung von Korpora mit Metadaten kann eine Recherche, deren Identifizierung und Wiederverwendung ermöglichen (Rümpel 2011: 31). Die Ergebnisse dieser Arbeit können also einerseits das eigene Forschungsdatenmanagement für Korpora unterstützen und andererseits deren Wiederverwendung durch eine auf Metadaten basierende Erschließung ermöglichen.

Nehmen wir die oben kurz skizzierten Erschließungswege aus Abbildung 1.1 als ersten Ansatz, dann entstünde ein n-zu-m-Verhältnis zwischen den einzelnen Erschließungen von Korpora zwischen und innerhalb von den Fächern, so dass sich jedes

[16] Neben Förderern wie der Deutschen Forschungsgemeinschaft (DFG) stellen auch Universitäten Anforderungen an Forschungsdatenmanagment, wie es beispielsweise auch die Humboldt-Universität zu Berlin in ihrem Grundsatzpapier beschreibt (Deutsche Forschungsgemeinschaft 2009; Humboldt-Universität zu Berlin 2014).

Fach einzeln alle potenziellen Korpora inner- und überfachlich erschließen müsste.[17] Um dies zu vermeiden, müsste es einerseits einen gemeinsamen und einheitlichen Zugriff und ein eben solches Verständnis von historischen Korpora geben (Abbildung 1.2). Korpora werden nicht mehr über die fachspezifischen Zugänge (Abbildung 1.1) erschlossen, sondern fachunabhängig. Andererseits kann nur eine *einheitliche* Beschreibung dieser verschiedenen Korpora deren Erschließung unterstützen. Eine solche einheitliche Erschließung soll, wie die vorliegende Arbeit zeigen wird, mit Hilfe einer abstrahierten, datenbezogenen, aber nicht fachbezogenen Dokumentation von Korpora möglich werden.

Wenn Korpora auf eine einheitliche Weise mit Metadaten dokumentiert werden können, dann können diese Metadaten wiederum als Grundlage für eine Metadatensuche in einem einheitlichen Zugang für Korpora dienen.[18] Das Finden von Korpora ist eine weitere Voraussetzung für deren Wiederverwendung (vgl. Abbildung 1.2). Gerade für die Suche nach semantischen Konzepten ist die computergestützte Verarbeitung und Dokumentation schwierig:

> Je nachdem, was Sie suchen [...] haben Sie ja ganz bestimmte Feinheiten, wo Sie hingucken, und das dem Computer beizubringen, ist ein extrem schwieriges Thema. (Keim 2016)

Solche Feinheiten können im Fall der historischen Korpora verschiedene Konzepte und Definitionen zu `Primärtext` darstellen. Wenn man die oben genannten Beispielkorpora betrachtet, dann können Primärtexte die Textes der historischen Zeitungen selbst sein, deren Digitalisate oder ihre korpuslinguistische Aufbereitung.[19] Dies dann in eine computergestützte Dokumentation für Korpora umzusetzen, so dass überfachlich auf ein semantisches Konzept von einem Primär- und Sekundärtextbegriff Bezug genommen werden kann, erscheint kaum umsetzbar und wenig zielführend.

Nehmen wir Abbildung 1.2 als Ausgangslage für diese Arbeit, dann muss eine für Forscherinnen und Forscher unbekannte Menge an unterschiedlichen Korpora so beschrieben werden, dass aus dieser Menge eine Auswahl auf Grundlage der vorab gegeben Informationen (Metadaten) unter einer eigenen Zielvorstellung möglich wird.

[17] Mit dem „Fach“ sind hier die Forscherinnen und Forscher gemeint.

[18] Diese Arbeit fokussiert sich auf die Entwicklung der Metadaten. Daneben ist in diesem Kontext auch die Entwicklung von Suchwerkzeugen, Speichersystemen und Repositorien wesentlich, können aber in dieser Arbeit nicht weiter diskutiert werden.

[19] Wie unterschiedlich dies diskutiert werden kann, zeigt Abschnitt 2.7.1.

Die Korpora werden nicht primär nach ihrem Fach sortiert und beschrieben, sondern nach ihren technisch-abstrakten Eigenschaften, die sie untereinander vergleichbar machen. Auf Grundlage dieser Eigenschaftseinordnungen können dann Korpora ganz unterschiedlicher Fachausrichtung und Architektur einheitlich erschlossen werden.

Über eine solche einheitliche Beschreibung durch Metadaten soll auf eine Menge an Korpora konzeptionell zugegriffen werden, ganz ähnlich wie die Suche nach Büchern über einen Online Public Access Catalogue (OPAC) einer Universitätsbibliothek. Durch die Suche mit einem OPAC muss nicht jedes Buch physisch in einer zu besuchenden Bibliothek durch das Lesen des Titels, des Fließtexts oder des Inhaltsverzeichnis nach den eigenen Suchkriterien überprüfen werden. Dieses Vorgehen wäre vergleichbar mit der Erschließung über die Ressource und ihr Format mittels eines Tools. Die strukturierte Suche mit dem OPAC ermöglicht es, nach relevanten Eigenschaften (Metadaten) von Büchern wie z. B. einem *Buchtitel* oder dem *Erscheinungsjahr* in einer oder mehreren Bibliotheken zu suchen. Die zentrale Frage dabei ist, welche Eigenschaften von Büchern für einen solchen Zweck und Zugang relevant sind.

Die gleiche Frage stellt sich auch für Korpora. Metadaten können im Prinzip von jedem über alles Mögliche erstellt werden (Hunter 2003). Eine klare Definition von Umfang und Zweck der Metadaten über einen definierten Typ an Forschungsdaten ist daher essenziell. Für das Bücher-Beispiel ist der wesentliche Zweck das Finden von unterschiedlichen Publikationsformen über deren Eigenschaften. Dies wird mit Hilfe von bestimmten Eigenschaften wie Titel oder Autor nicht aber mit der Farbe des Covers umgesetzt. Die Farbe des Covers ist eine Eigenschaft, die für den vorgebenden Zweck genutzt wird. Ein OPAC enthält daher Metadaten, die als Suchkriterien fungieren, die nur für einen bestimmten Zweck relevante Eigenschaften von Büchern tragen. Wenn sich der Kontext von einer traditionellen Bibliothek – ein Haus mit Büchern – hin zu einer digitalen Bibliothek ändert, aber sehr ähnliche Aufgabe für digitale Bücher erfüllt werden sollen, zeigt sich dies auch Änderungen der Anforderungen an die Metadaten.[20] Auf den Webseiten von Metadata Encoding and Transmission Standard (METS) wird dieser Zusammenhang so erklärt:

> Wenn eine Bibliothek Metadaten zu einem Buch in ihrem Bestand erfasst, wird dieses Buch nicht in eine Reihe einzelner Blätter zerfallen,

[20] Vgl. für den Aufbau und die Funktionsweise von digitalen Bibliotheken z. B. Solodovnik (2011) und Xie und Matusiak (2016).

> weil keine Strukturangaben über die innere Ordnung des Buches erhoben werden. Noch werden Forscher das Buch schlechter nutzen können, wenn nicht angegeben wurde, dass es mit einer Ryobi Druckmaschine hergestellt wurde. Gleiches gilt jedoch nicht für die digitale Version desselben Buches. Ohne Metadaten zur Struktur sind die Seitenabbildungen oder die Textdateien, aus denen es besteht, so gut wie wertlos. Und ohne technische Metadaten über den Digitalisierungsprozess können Leser nicht sicher sein, wie genau die digitale Version die ursprüngliche Vorlage wiedergibt.[21]

Hier wird deutlich herausgearbeitet, dass Metadaten verschiedene Dinge für einen vergleichbaren oder unterschiedlichen Zweck beschreiben können. Die Auswahl der Metadaten hängt von den vorhandenen Eigenschaften des zu beschreibenden Objektes selbst und von der Verwendung dieses Objektes ab. Übertragen auf historische Korpora heißt das, Metadaten müssen diese so beschreiben, dass alle relevanten Informationen für eine Wiederverwendung des Korpus vorliegen. Dabei besteht auch hier ein Unterschied zwischen der historischen Vorlage (Text) und dem Korpus, das ein Digitalisat des Textes beinhaltet. Zentrale Fragen für diese Arbeit sind daher:

- Welche gemeinsamen Eigenschaften besitzen historische Textkorpora?
- Welche Wiederverwendungsszenarien für historische Korpora gibt es?
- Welche Eigenschaften sind für die Erschließung von historischen Korpora relevant und müssen für den Zweck der Wiederverwendung dokumentiert werden?
- Wie können diese Eigenschaften als Metadaten repräsentiert werden?
- Wie kann über diese Metadaten weiter abstrahiert werden, um ein allgemeines Beschreibungsmodell für historische Korpora zu entwickeln?

1.2 Methodik und Aufbau der Arbeit

Die vorliegende Arbeit befasst sich mit der Dokumentation von historischen Korpora. In Kapitel 2 werden die Eigenschaften von historischen Korpora als ein spezieller Typ von Korpus vorgestellt. Dabei wird besonders auf die Beziehung zwischen der historischen Vorlage (Text) und ihrer digitalisierten Form im Korpus eingegangen.

[21] `http://www.loc.gov/standards/mets/METSOverview.v2_de.html` (besucht am 16.09.2016).

Diese Arbeit stützt sich dabei auf eine Menge von historischen Korpora, die die empirische Grundlage dieser Arbeit bilden. Als durchgängiges Beispiel wird in dieser Arbeit das Register in Diachronic German Science-Korpus (RIDGES) (Lüdeling et al. 2014; Odebrecht et al. 2017) verwendet. So werden die Eigenschaften anhand mehrerer authentischer Beispiele herausgearbeitet. In Kapitel 3 werden darauf aufbauend Wiederverwendungsszenarien für die historischen Korpora erarbeitet. Zentrale Aspekte für diese Arbeit sind weiterhin die Einordnung und Funktion der Metadaten. Daher muss der Begriff `Metadaten` im wissenschaftlichen Kontext verortet und dann in Beziehung zu den hier untersuchten Textkorpora gestellt werden. Dazu wird aufgezeigt, welche Arten von Metadaten mit welchen Funktionen und Strukturen für Forschungsdaten allgemein aus der Perspektive der Informations- und Bibliothekswissenschaften bereits etabliert sind oder zumindest Verwendung finden sowie für die Beschreibung von historischen Korpora eingesetzt werden können (Kapitel 4).

Auf diese Weise werden mithilfe der in Kapitel 2, Kapitel 3 und Kapitel 4 erarbeiteten Voraussetzungen die Anforderungen an eine Korpusdokumentation für historische Korpora herausgearbeitet. Die Arbeit wird die bisherigen Ansätze der Metadatenschemata, die für Textkorpora genutzt werden können, vgl. Dublin Core (DC), ISLE Meta Data Initiative (IMDI), Component MetaData Infrastructure (CMDI), Metadata Encoding and Transmission Standard (METS) und Text Encoding Initiative (TEI), vor dem Hintergrund der vorher definierten Anforderungen diskutieren (Kapitel 5).

Die Aufgabe eines solchen Metadatenschemas ist es, eine gemeinsame einheitliche Beschreibungsebene zu liefern, die alle relevanten Eigenschaften der historischen Korpora abbildet. Da eine solche Abstrahierung von allen relevanten Eigenschaften historischer Korpora bislang in den bisherigen Ansätzen nicht oder nur teilweise zugrunde gelegt wird, wird eine solche Abstrahierung in dieser Arbeit mit einer Drei-Ebenen-Modellarchitektur und dem Metamodell für Korpusmetadaten Metamodell für Korpusmetadaten (MKM) erarbeitet (Kapitel 6). Somit befasst sich diese Arbeit theoretisch mit der Modellierung von Metadaten für den Forschungsdatentyp `historisches Textkorpus` und nutzt dabei eine formale Modellierungssprache. Anschließend wird ein Vorschlag für die Realisierung des MKM mit Hilfe der TEI gemacht, mit dem dann Korpusmetadaten in Anwendungen ausgelesen, indexiert und angezeigt werden können (Kapitel 7).

2 Korpora

In diesem Kapitel werden historische Korpora definiert und eingeordnet sowie deren Eigenschaften beschrieben, die Forscherinnen und Forscher kennen müssen, um Korpora (wieder-)verwenden zu können.

Ganz allgemein sind Korpora digitale Forschungsdaten, mit denen zwei wichtige Aspekte verbunden sind, der `Forschungsprozess` und der `Forschungsdatenzyklus`.

> Unter digitalen Forschungsdaten verstehen wir dabei alle digital vorliegenden Daten, die während des Forschungsprozesses entstehen oder ihr Ergebnis sind. Der Forschungsprozess umfasst dabei den gesamten Kreislauf von der Forschungsdatengenerierung, z. B. durch ein Experiment in den Naturwissenschaften, eine dokumentierte Beobachtung in einer Kulturwissenschaft oder eine empirische Studie in den Sozialwissenschaften, über die Bearbeitung und Analyse bis hin zur Publikation und Archivierung von Forschungsdaten. (Kindling und Schirmbacher 2013: 130)

Korpora sind demnach Daten, die in einem Forschungsprozess erzeugt werden. So können Forschungsdaten in Abhängigkeit von dem jeweiligen Forschungsprozess als digitales Artefakt, als prozessierbare Information oder als interpretierbarer Text verwendet werden (Owens 2011). Korpora können ein Produkt des Forschungsprozesses selbst oder ein Beiprodukt dieses Prozesses darstellen. Deren Metadaten können entweder als eine Art „Überbleibsel“ eines Vorgangs verstanden oder als Input für beispielsweise eine Suchsoftware für Facettensuchen oder als Grundlage für eine Dokumentation verwendet werden (Kapitel 4).

Forschungsdaten wird eine Art Lebenszyklus zugeschrieben, der wie der Forschungsprozess konstituierend für Forschungsdaten ist.[22] Unter dem `Lebenszyklus von Forschungsdaten` (Digital Curation Centre 2010) wird allgemein jeder Schritt des Forschungsdatenmanagements verstanden, der die Idee, das Forschungsdatendesign, dessen Umsetzung, Auswertung und Veröffentlichung beinhaltet. Jeder Typ

[22] Verschiedene Wissenschaften können hingegen ganz unterschiedliche Konzepte von Forschungsdaten besitzen (Büttner et al. 2011: 15).

von Forschungsdatum (und damit auch Korpora) und jede Art der Bearbeitung kann folglich in den Lebenszyklus von Forschungsdaten eingeordnet werden (Rümpel 2011: 27). Korpora können also in Abhängigkeit vom einem Punkt oder einer Sequenz ihres Lebenszyklus definiert werden. Dabei ist der konkrete Forschungsdatenzyklus meist fachbezogen und forschungsdatenspezifisch. So wird auf den Webseiten von DARIAH in diesem Zusammenhang festgestellt,

> 1. dass es keinen allgemeingültigen einheitlichen Begriff eines Forschungsdatenzyklus gibt,
> 2. dass die Ansätze in ihrer Granularität stark variieren und
> 3. dass die einzelnen Ausprägungen entscheidend von der Fachdisziplin geprägt werden, aus der ein Ansatz stammt.[23]

Ein wichtiger Aspekt für die Definition von Forschungsdaten ist damit, dass sie relativ zu ihrem Fach und damit zur Forschungsfrage (Forschungsprozess) sowie zu ihrem Zustand (Lebenszykus) beschrieben und definiert werden. Dies wird auch bei der Unterscheidung zwischen Primärdaten und Sekundärdaten deutlich:

> Dieser Begriff „Primärdaten“ sorgt immer wieder für Diskussion, denn die Definition des Begriffs ist sehr von der eigenen Rolle in der wissenschaftlichen Wertschöpfungskette bestimmt. Für den einen sind „Primärdaten“ der Datenstrom aus einem Gerät, z. B. einem Satelliten. In der Fernerkundung werden diese Daten „Level 0“ Produkte genannt. Für andere sind „Primärdaten“ zur Nachnutzung aufbereitete Daten, ohne weiterführende Prozessierungsschritte. Wieder andere differenzieren nicht nach Grad der Verarbeitung sondern betrachten alle Daten, die Grundlage einer wissenschaftlichen Veröffentlichung waren, als Primärdaten. Der begrifflichen Klarheit wegen sollte daher das Präfix „Primär-“ nicht mehr verwendet werden und statt dessen nur noch von wissenschaftlichen Daten oder Forschungsdaten gesprochen werden. (Klump 2009: 104-105)

Wenn der Forschungsdatenzyklus jeweils spezifisch für einen Typ von Forschungsdaten ist, dann ist auch eine allgemeine Definition von `Primärdatum` für Korpora nicht ohne Weiteres möglich. Beispielsweise können die Ergebnisse der Forschung an einem bestimmten Forschungsdatum in Form von wissenschaftlichen Artikeln publiziert werden. Diese wissenschaftlichen Artikel, die als Ergebnis eines Forschungsprozesses verstanden werden können, stellen beispielsweise in Conrad (1996) die

[23] `https://de.dariah.eu/bestehende-konzepte` (besucht am 31.12.2016).

Forschungsgrundlage, deren Ergebnis wiederum in einem wissenschaftlichen Artikel publiziert wird. Forschungsdaten stehen damit und nach obiger Definition in Relation zur Forschungsfrage. Der wissenschaftliche Artikel von Conrad (1996) ist dann ein Produkt des Forschungsprozesses. Ein Korpus aus wissenschaftlichen Artikeln ist ebenfalls ein Produkt des Forschungsprozesses. Im Sinne des Forschungsdatenzyklus haben beide Produkte verschiedene Stadien wie Konzeption, Erstellung oder Korrektur durchlaufen und sind gespeichert und zugänglich. Auf der Basis eines Produktes des Forschungsprozesses (wissenschaftliche Artikel) arbeitet ein weiterer Forschungsprozess, der wiederum ein Produkt (Korpus) erzeugt.[24]

Den Forschungsprozess und den Lebenszyklus der Forschungsdaten, der ersteren begleitet, beschreiben Forschungsdaten als Teil eines Prozesses oder als dessen Produkt. Eine wichtige Schlussfolgerung daraus ist, dass bei der Beschreibung der Forschungsdaten neben einer technischen und funktionalen auch eine zeitliche Perspektive mit einbezogen werden muss: Soll der laufende Prozess oder das Produkt am Ende eines Prozesses beschrieben werden? Diese Frage wird in Abschnitt 4.4 bei der Klassifikation der Metadaten wieder aufgegriffen.

Aus den obigen Überlegungen werden folgende Aspekte in dieser Arbeit berücksichtigt, mit denen Korpora als eine spezielle Art von digitalen Forschungsdaten beschrieben werden müssen:

I ein Produkt des Forschungsprozesses

II ein Produkt des Lebenszyklus

In dieser Arbeit wird also mit dem Begriff `Korpus` auf das Produkt eines Forschungsprozesses und eines Forschungsdatenzyklus referiert. Korpora werden als ein gespeichertes, in diesem Sinne nicht flüchtiges, zugängliches Produkt verstanden. Analytische Erzeugnisse wie wissenschaftliche Artikel werden dabei nicht mehr berücksichtigt (Abschnitt 2.7.3 und Kapitel 3).

In Abschnitt 2.1 wird eine korpuslinguistische Definition von Korpora gegeben. Danach werden die Eigenschaften des Korpustyps *historisches Korpus* im Detail vorgestellt (Abschnitt 2.7). Beide Abschnitte diskutieren und erklären Korpora auf einer beschreibenden Ebene: Eigenschaften der Korpora bezüglich ihrer Korpustypen, Annotationskonzepte, Metadaten, Korpusarchitektur und -verarbeitung werden nur skizziert, deren konkrete technische Umsetzungen nicht vollständig thematisiert.

[24] In Abschnitt 4.4 wird noch einmal ausführlich der zeitliche Bezug und der Begriff `Produkt` in diesem Zusammenhang diskutiert.

Ziel ist es, Kategorien und Beschreibungsmerkmale von Korpora zu identifizieren, die dann in die Modellierung einfließen können.

2.1 Definition von Korpus

In der Korpuslinguistik wird ein `Korpus` als eine Sammlung digitalisierter natürlich-sprachlicher Äußerungen wie Texte, Transkripte, oder Audioaufnahmen verstanden. Diese Sammlung kann mit weiteren Interpretationen in Form von `Annotationen` angereichert werden (Kuebler und Zinsmeister 2015; McEnery und Hardie 2012). Der Begriff `Korpus` ist in den verschiedenen Teildisziplinen der Linguistik, Korpuslinguistik und Computerlinguistik bereits etabliert (Lüdeling und Zeldes 2007: 149) und wird auch in anderen Fächern genutzt (vgl. Romary 2013). Nicht in dieser Arbeit berücksichtigt werden elektronische Lexika und Sprachatlanten. Diese wären auch digitale Forschungsdaten nach obigen Definitionen, sie beinhalten jedoch selten authentisches sprachliches Material in einem Kontext.

`Annotationen` sind explizite Zuweisungen von Kategorien (Tags) zu einem Token oder eine Sequenz von Token (Odebrecht et al. 2017). Weitere Informationen über das Korpus wie z. B. die Datenquelle und Auswahlkriterien befinden sich in den `Metadaten`.

Die Forschungsfrage kann bestimmen, welche natürlich-sprachlichen Äußerungen für ein Korpus gewählt werden. Damit kann sie einen Einfluss auf das `Korpusdesgin` haben.

> Corpus design, that is, how much of what kinds of texts are included, determines to a certain extent how a corpus can be used, especially if one wants to make quantitative statements. But even if a corpus is used merely as an 'example bank', its design may be relevant because given structures and contexts will only be found in certain corpus types. (Lüdeling et al. 2016: 601)

Die Art der sprachlichen Äußerungen, die im Rahmen des Korpusdesigns ausgewählt werden, bestimmt damit den Typ des Korpus. Ein `Korpustyp` kann eine oder mehrere Datengrundlagen in Form von beispielsweise Text, Ton, Bild oder Video besitzen (vgl. Abschnitt 2.2). Das Korpusdesign und der Korpustyp ermöglichen und limitieren die Art der Forschung, die mit einem Korpus gemacht werden kann.

Wie mit Hilfe von Korpora in der Linguistik geforscht wird, kann grob in zwei Arten unterteilt werden, die korpusbasierte und die korpusgetriebene Forschung (Lüde-

ling und Zeldes 2007; McEnery und Hardie 2012). Beide Forschungsrichtungen sind weit über die Linguistik hinaus etabliert. Ein Beispiel für korpusgetriebene Untersuchungen sind Kollokationsanalysen (n-gram-Analysen, vgl. bereits Sinclair 1995), die Vorkommen fester Abfolgen von z. B. Wortformen in einem Korpus (technisch ein n-gram) untersuchen (vgl. Biber und Conrad 1999). Dabei wird das Korpus selbst genutzt, um Hypothesen über Sprache zu generieren. Eine korpusbasierte Studie erarbeitet Hypothesen über ein sprachliches Phänomen, das empirisch mit Hilfe von einem Korpus überprüft wird, wie z. B. die Untersuchung der Kasusmarkierungen in Präpositionalphrasen mit der Wortform *voller* im Deutschen (Zeldes erscheint). Neben dieser Unterscheidung kann eine Analyse auch qualitativ und/oder quantitativ durchgeführt werden (Lemnitzer und Zinsmeister 2006). Korpusgetriebene – meist quantitative – Forschung kann dann ein anderes Korpusdesign als korpusbasierte Forschung benötigen.

Außerhalb der Linguistik wird ein weiteres methodische Spektrum mit der Dichotomie zwischen `close reading` und `distant reading` beschrieben. `Distant reading` (Moretti 2007; Gooding et al. 2013) beschreibt eine quantitative Analysemethode auf der Grundlage einer Vielzahl an (digitalen) Texten. Dabei müssen die einzelnen Texte oder Abschnitte der Texte nicht gelesen und verstanden werden, wie es das Ziel des `close readings` ist (vgl. für eine Einführung Simanowski 2011; Federico 2015). Dieses Spektrum ist vergleichbar mit den methodischen Einordnungen, die in der Korpuslinguistik getroffen werden.

Das Korpus liegt dazu bereits vor und ist ein gespeichertes, zugängliches Produkt von mehreren Bearbeitungsschritten (Lebenszyklus). Für die korpusbasierte und korpusgetriebene Methode wie für die close- und distant-reading-Methoden sind Korpora die Analysegrundlage. Korpora sind damit ein Produkt des Forschungsprozesses unabhängig von der jeweiligen Analysemethode.

Wenn Korpora Sammlungen authentischen sprachlichen Materials sind, die mit Annotationen versehen werden können (aber nicht müssen) und als Analysegrundlage für ein Forschungsvorhaben ausgewählt werden können, dann sind digitale Textsammlungen (auch *Collections*/*Kollektionen* genannt) ebenfalls Korpora. Das können z. B. die digitale Textsammlung der Projekte *Gutenberg*[25] oder Sammlungen digitaler Protokolle sein, wie sie der Deutsche Bundestag[26] veröffentlicht. Diese Textsammlungen sind nicht für eine spezifische Forschungsfrage erstellt worden und sind

[25] Digitalisierte gemeinfreie Texte wie Märchen oder klassische Literatur vgl. `https://www.gutenberg.org` (besucht am 27.01.2017).

[26] Protokolle der Parlamentsreden `https://www.bundestag.de/protokolle` (besucht am 27.01.2017).

nicht zwingend mit weiteren Annotationen versehen, können aber die empirische Grundlage für die Beantwortung einer Forschungsfrage stellen.[27] Damit wird hier auch keine Unterscheidung zwischen Archiven, Kollektionen und Korpora gemacht, wie sie Wegera (2013) funktional trifft: Korpora seien zweckgebunden, Archive besäßen keinen eindeutig bestimmten Zweck. Eine solche funktionale Unterscheidung zwischen Textarchiven, Textsammlungen und Korpora verfolge ich in dieser Arbeit nicht, weil die Wiederverwendung eines Korpus oder eines Archivs eine Änderung des initialen Zwecks der Erstellung darstellt. Unter einem solchen Zweck verstehe ich in diesem Zusammenhang z. B. auch die Beantwortung einer Forschungsfrage, wofür ein Korpus (wieder-)verwendet werden kann. Damit werden hier beispielsweise das Bonner Frühneuhochdeutschkorpus[28] (Solms und Wegera 1998) und das Deutsches Textarchiv (DTA)[29] (Geyken 2013) als Korpora definiert.

Der Forschungszweck beziehungsweise eher die konkreten Forschungsfragen, die an ein Korpus gestellt werden können, werden daher nicht als Kriterium zu deren Klassifikation hinzugezogen.

Der linguistische Begriff `Korpus` lässt sich mit dem hier vorgestellten, allgemeinen Begriff von `digitalen Forschungsdaten` verbinden, erhält damit zusätzlich noch die Beschreibungskomponenten des Forschungsdatenzyklus und des Forschungsprozesses und besitzt somit einen überfachlichen Bezug.

Der Korpustyp, die Annotationskonzepte, die Annotationen und die daraus resultierende Korpusarchitektur bestimmen maßgeblich die Eigenschaften von Korpora, weshalb in den nachfolgenden Abschnitten diese Begriffe näher erläutert werden. Daran anschließend werden die Besonderheiten der Korpusarchitekturen historischer Korpora diskutiert (Abschnitt 2.7).

2.2 Korpustyp

Die Auswahl der sprachlichen Äußerung bestimmt den Korpustyp: `Textkorpora` (Hundt 2008), die genuin geschriebene Sprache oder Transkripte gesprochener Sprache beinhalten, `Korpora gesprochener Sprache` oder `speaking corpora` (Ballier und P. Martin 2015; Wichmann 2008), die gesprochene Sprache in Form von Audioaufnahmen beinhalten, oder `multimodale Korpora` (Allwood 2008), die beispielsweise zusätzlich Audio-, Bild- oder Videomaterial beinhalten. Diese Unterscheidun-

[27] Vgl. z. B. Kytö (2011) für einen Überblick zu englischen Textsammlungen.
[28] `https://korpora.zim.uni-duisburg-essen.de/fnhd/` (besucht am 27.01.2017).
[29] `http://www.deutschestextarchiv.de/` (besucht am 27.01.2017).

gen basieren auf der linguistischen Einordnung von Modalitäten wie Mündlichkeit und Schriftlichkeit:

> The terms 'written' and 'spoken' are normally taken to refer to the (primary) channel of transmission: texts can be transmitted in the written or spoken medium. But they can also be written to be spoken (for example lectures, political speeches, some kinds of radio broadcasts) or they can be transcribed speech (i. e. medially 'written' recordings of originally 'spoken' language). Therefore, in addition to the medial aspect (i. e. the channel of transmission), a distinction has to be made between conceptually 'literal' and 'oral' texts. Both aspects – the medial and the conceptual – overlap. (Hundt 2008: 169)

So ist ein Transkript gesprochener Sprache medial schriftlich, im Vergleich zu der medial mündlichen Audioaufnahme derselben sprachlichen Äußerung. Konzeptionell hingegen ist das Transkript mündlich. Ein Transkript ist damit kein typischer Vertreter von medialer Schriftlichkeit (vgl. auch Koch und Österreicher 1985).

Neben diesen Einteilungen diskutieren Ágel und Hennig (2006) in diesem Zusammenhang die Begriffe *Nähe und Distanz von Sprache*. Ihre Modellierung bezieht sich dabei auf die Kommunikationsbedingungen, „die *nachweislich* für das Vorhandensein oder Nichtvorhandensein bestimmter grammatischer Merkmale verantwortlich sind“ (Ágel und Hennig 2006: 24). Die Kommunikationsbedingungen stellen also ein weiteres Beschreibungsmerkmal von Äußerungen. So kann bei der Transkription von historischen Texten wie z. B. Predigten im Rahmen einer Korpuserstellung folgende Frage gestellt werden: Inwieweit wird ein medial mündlicher oder schriftlicher, nähesprachlicher oder distanzsprachlicher Text transkribiert? In diesem Kontext sind die Begriffe `Primärtext` und `Text` für historische Korpora mit ganz unterschiedlichen Auffassungen verbunden, sodass die konzeptionelle Einordnung unter verschiedenen Aspekten betrachtet werden muss (vgl. hierzu Abschnitt 2.7.1).

Diese verschiedenen Korpustypen unterscheiden sich zumindest grundlegend nach ihrem Modus der Sprache sowie nach der Form, wie sie abgelegt sind.

Jede dieser sprachlichen Ressourcen können homogen, balanciert, heterogen oder opportunistisch zusammengestellt werden.[30] Ein häufiger Typ der homogen gesampelten Textkorpora sind Zeitungskorpora, die Artikel, Kommentare, Abschnitte oder

[30] Vgl. für einen ersten Überblick zum Korpusdesign Hunston (2008). Darüber hinaus ist die Entscheidung, welche Auswahl als repräsentativ für Sprache motiviert werden kann, schwer zu treffen (vgl. Z. B. Biber 1993).

Werbetexte aus Zeitungen enthalten, siehe z. B. Telljohann et al. (2003) für Korpora mit moderner Zeitungssprache und Demske (2007) für Korpora historischer Zeitungssprache des Deutschen. Das sprachliche Material kann ursprünglich sowohl analog wie im Fall des historischen Zeitungskorpus oder auch digital wie im Fall des modernen Zeitungskorpus vorliegen. Ein Beispiel für opportunistisch gesampelte Textkorpora sind Web-Korpora, die aus einer Vielzahl an automatisch im Web gesammelten Texten bestehen und ursprünglich bereits digital vorliegen, wie DEUTSCHES WEB ALS CORPUS (deWaC) (Baroni et al. 2009)[31] oder CORPORA FROM THE WEB (COW) (Schäfer und Bildhauer 2012)[32].

Ein Beispiel für balancierte Korpora, die auf Transkripten von geschriebenen sprachlichen Ressourcen basieren, sind Lernerkorpora wie das FEHLERANNOTIERTE LERNERKORPUS (Falko)[33]. Es besteht aus Transkripten von handschriftlich verfassten Aufsätzen und aus direkt digital erstellten Essays von sowohl Muttersprachlern und als auch Lernern, die Deutsch als Fremdsprache lernen (Reznicek et al. 2012). Korpora können ebenfalls aus Transkripten von historischen Quellen toter Sprachen wie des Koptischen (Schroeder et al. 2016; Zeldes und Schroeder 2015) oder des Altäthiopischen (Vertan et al. 2016)[34] bestehen. Korpora der gesprochenen Sprache wie das GESPRÄCHSCORPUS (GECO) (Schweitzer und Lewandowski 2010, 2013) beinhalten Audioaufnahmen von freien Dialogen, die als Tonspur und Transkription im Korpus enthalten sind. Multimodale Korpora wie das BERLIN MAP TASK CORPUS (BeMaTaC) (Sauer und Lüdeling 2016)[35] besitzen Video- und Tonspuren des sprachlichen Materials.

Korpora, die für die Untersuchung von gesprochener Sprache erstellt werden, können aus zwei Perspektiven betrachtet werden. Ein Korpus ohne Audioaufnahmen besteht technisch gesehen aus Transkripten, also Texten. Konzeptionell, auch aus fachlicher Perspektive, ist das Gesprochene und damit das Audiosignal primär. Das Primärdatum ist einerseits als theoretische Kategorie, ähnlich den Beispielen zu Wortarten und Autoren (Abschnitt 2.3), zu verstehen, muss aber auch in einem konkreten Format abgebildet werden können. So kann es bei einer korpuslinguistischen Einteilung zwei Perspektiven geben: eine konzeptionell-fachbezogene Perspektive und eine technische Perspektive.

[31] `https://www.sketchengine.co.uk/xdocumentation/wiki/Corpora/DeWaC` (besucht am 27.01.2017).
[32] `http://hpsg.fu-berlin.de/cow/` (besucht am 27.01.2017).
[33] `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/standardseite` (besucht am 10.11.2017)
[34] `https://www.traces.uni-hamburg.de/` (besucht am 27.01.2017).
[35] `http://u.hu-berlin.de/bematac` (besucht am 27.01.2017)

Ähnliche Fragen stellen sich auch für historische Korpora (vgl. Himmelmann 2012): Ist beispielsweise bei einem historischen Korpus das historische Buch, das Foto des Faksimiles oder das Transkript Primärtext? Für eine fachübergreifende Dokumentation von Korpora ist dies entscheidend. Aus welcher Perspektive soll dokumentiert werden? Wie muss der Modus der authentischen sprachlichen Äußerung und der Datengrundlage des Korpus aufgegriffen werden? Wie kann der Text innerhalb des Korpus beschrieben werden? Diese Aspekte werden mit dem Fokus auf textbasierte Korpora näher in Abschnitt 2.7 und in Kapitel 6 diskutiert.

Relativ klar ist die Abgrenzung in Bezug auf die Datengrundlage eines Korpus, wenn man die Korpustypen aus einer eher technisch-medialen Perspektive hinsichtlich ihres Formats wie Text, Bild, Ton und Video unterteilt. In dieser Arbeit werden Korpora berücksichtigt, die keine Audio- oder Videodateien besitzen. Eine Unterscheidung zwischen konzeptioneller Mündlichkeit und Schriftlichkeit bzw. eine Unterscheidung zwischen Nähe- und Distanzsprache wird in dieser Arbeit mit der Definition des Korpustyps nicht getroffen. Damit geht es in dieser Arbeit ausschließlich um den Korpustyp `Textkorpus`.

Entscheidend ist, nur die Datengrundlage und damit den Korpustyp und keine weiteren Konzepte zu erfassen, damit die jeweils fachspezifischen Forschungsfragen und Ziele nicht in den Fokus der Korpusdokumentation rücken. Damit muss für die Datengrundlage keine Erweiterung der linguistischen Definition von *Korpus* erfolgen, selbst wenn diese auch auf nicht ausschließlich linguistische Datensätze angewandt wird.

2.3 Kategorisierungen und Annotationsrichtlinien

Annotation werden als die Zuweisung von Kategorien zu Exponenten (Abschnitt 2.1) werden und sind sie immer auch Interpretationen. Ihre Bedeutung und deren Zuweisung sowie Auswertung kann nicht komplett unabhängig von der jeweiligen Forschungsfrage oder dem Forschungskontext getrennt werden (Lüdeling 2011). Annotationen eines Korpus können ebenfalls für andere Forschungsvorhaben wiederverwendet werden. Annotationen sind immer stark auf den Forschungsprozess bezogen, werden aber nicht immer nur durch einen einzigen Forschungsprozess (Forschungsfrage) definiert. Im Prinzip können alle theoretisch möglichen Kategorien zu den unterschiedlichsten Exponenten zugewiesen werden.

Annotationen können in einem Korpus flach, hierarchisch und diskontinuierlich verweisend zugeordnet sein (Abschnitt 2.4.2). Typischerweise werden Annotationen

mit Hilfe von Annotationsschemata oder -guidelines manuell, semi-automatisch oder automatisch erstellt (Leech 1993). Diese Annotationsschemata enthalten alle für die jeweiligen Forschungsfragen relevanten Kategorien, deren Definitionen und eine Annotationsanleitung, wann und wie diese zugewiesen werden sollen (Kuebler und Zinsmeister 2015: 33-36). Wenn Annotationen Interpretationen und das Produkt eines Forschungsprozesses sind, dann können wenige feste, allgemein für verschiedene Korpustypen gültigen Annotationsstandards oder feste Annotationskonzepte innerhalb eines Fachs wie auch überfachlich identifiziert werden, weil es verschiedene, teilweise auch konfligierende Interpretationen desselben Sachverhalts geben kann oder entwickelt werden können. Dies soll anhand zweier Beispiele kurz belegt (Abschnitt 2.3.1 und Abschnitt 2.3.2) und im Weiteren berücksichtigt werden.

2.3.1 Beispiel für linguistische Annotationen

Es gibt zahlreiche Standardisierungsvorschläge für verschiedene linguistischen Domänen, z. B. für Syntax (Romary et al. 2015), für die morphosyntaktische Domäne (Romary und Witt 2012) und für mehrere linguistische Domänen (Ide und Suderman 2014).[36]

Ein Beispiel ist die Kategorie *Wortart*, die für viele linguistische Studien zentral ist. Sie kann abhängig vom Korpustyp und der Forschungsfrage unterschiedlich annotiert werden (vgl. für einen ersten Überblick Atwell 2008). Das Stuttgart-Tübingen-Tagset (STTS) für Wortarten (Schiller et al. 1999) ist ein häufig genutztes Tagset, das sich als eine Art Standard etabliert hat. Die Korpora RIDGES[37], Deutsch Diachron Digital – Referenzkorpus Altdeutsch (DDD-AHD) (Donhauser et al. 2014), das Referenzkorpus Frühneuhochdeutsch[38] und das Fürstinnenkorrespondenzkorpus(Lühr et al. 2014)[39] enthalten Wortartenannotationen, die diesen De-facto-Standard für ihre eigenen Forschungsfragen und historischen Texte anpassen und anwenden.[40] Alle genannten Korpora erstellen die Grundlage für

[36] Einen ersten Überblick über linguistische Annotationsformate geben Lehmberg und Wörner (2008).

[37] Annotationsrichtlinie unter `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/documentation/documentation-v4.1-de` (besucht am 02.02.2017).

[38] `http://www.ruhr-uni-bochum.de/wegera/ref/` (besucht am 08.08.2016), Annotationsrichtlinie nach Historisches Tagset (HiTS) (Dipper et al. 2013).

[39] Annotationsrichtlinie unter `http://dwee.eu/Rosemarie_Luehr/userfiles/downloads/Projekte/Dokumentation.pdf` (besucht am 02.02.2017).

[40] Für einen Vergleich zwischen verschiedenen Annotationsrichtlinien für Wortarten vgl. Kuebler und Zinsmeister (2015: 50-54).

korpuslinguistische Untersuchungen von historischen Sprachstufen des Deutschen.

Ein Vergleich der jeweiligen Annotationsschemata für Wortarten, basierend auf dem STTS, zeigt, wie unterschiedlich diese Schemata angepasst, umgesetzt und genutzt werden. Tabelle 2.1 zeigt dies anhand der Kategorien für Adjektive.

Tabelle 2.1: *Vergleich von Wortartenannotationen für Adjektive. Annotationskategorien basierend auf dem STTS.*

RIDGES	**DDD-AHD**	**Fürstinnen**	**HiTS**	**Beschreibung**
ADJA	ADJ	ADJA	ADJA	attributives Adjektiv
ADJD	ADJD	ADJD	ADJD	adverbiales oder prädikatives Adjektiv
	ADJE			Adjektiv, attributiv, Teil eines Eigennamens
	ADJN		ADJN	Adjektiv, attributiv, nachgestellt
	ADJNE			Adjektiv, attributiv, nachgestellt, Teil eines Eigennamens
	ADJO			Adjektiv, ordinal, attributiv
	ADJON			Adjektiv, ordinal, attributiv, nachgestellt
	ADJOS			Adjektiv, ordinal, substantiviert
	ADJS			Adjektiv, substantiviert
		ADJAA		Attributives Adjektiv, abgekürzt
		ADJDA		Adverbiales oder prädikatives Adjektiv, abgekürzt
			ADJS	Adjektiv, substituierend

Das STTS nach Schiller et al. (1999) sieht zwei Tags für Adjektive vor: ADJD für adverbiales oder prädikatives Adjektiv und ADJA für attribuierende Adjektive. Für die Annotation von Adjektiven in diesen historischen Korpora werden neun, vier oder auch nur zwei Tags mit unterschiedlichen Kategorisierungskriterien, wie Funktion, Position oder Bezugswort, verwendet. In Abhängigkeit der Forschungsfrage, der Sprachstufe und der konkreten Aufbereitungsform sind demnach unterschiedliche Kategorisierungen und damit unterschiedliche Tags für die Wortartenannotation Part of Speech (pos) gewählt. Teilweise werden gleiche Tags – z. B. ADJS – für unterschiedliche Kategorien verwendet.

An diesem kleinen Beispiel zeigt sich, dass es selbst für eine elementare linguistische Kategorie wie Wortart (genauer: Adjektiv) verschiedene unterschiedlich motivierte Kategorisierungen gibt. Wenn in einer Korpusdokumentation allein vermerkt sein würde, dass Wortartenannotationen enthalten sind, dann kann damit auf sehr unterschiedliche Umsetzungen referiert werden. Auch eine Spezifikation wie *angelehnt an einen Standard wie dem STTS* wäre nicht ausreichend, da auch Standards

abgewandelt werden können, um bestimmte ggf. fehlende oder feinere Kategorien mit abbilden zu können. In beiden Fällen würde eine Korpusdokumentation für andere Forscherinnen und Forscher nicht genügend Informationen darüber enthalten, was genau annotiert wurde. Vielmehr müssten die jeweiligen Korpusdokumentationen genau spezifizieren, welche Konzepte und Kategorisierungen in diesem Fall für Wortartenannotationen (und für jede weitere Annotation im Korpus) verwendet werden (und welche nicht).

2.3.2 Beispiel für editorische Annotation

Ein nicht ausschließlich linguistisches Beispiel für De-facto-Standards und deren unterschiedliche Anwendung sind die TEI-Guidelines (TEI Consortium 2015), die eine Art Annotationsschema für die digitale Repräsentation von Texten darstellen. Beispielsweise umfassen die Annotationsrichtlinien die Ausweisung von graphischen Eigenschaften und das MarkUp von Texten wie Zeilenumbrüche, Überschriften oder Hervorhebungsarten. Neben Informationen zur Textgestaltung können beispielsweise auch Personen entweder in einem Text identifiziert und ausgewiesen oder als Metadatum angegeben werden. Drei passende Elemente aus diesen Guidelines sind `<bibl>`[41], `<author>`[42] und `<docauthor>`[43]. Die verschiedenen Elemente können zusätzlich auch Attribute erhalten und unstrukturierten Text in Form einer Zeichenkette beinhalten (vgl. Beispiele 1 bis 5).[44]

1. Beispiel mit TEI-Elementen für die einfache Auszeichnung von Autoren:

```
<bibl>Jules Verne, Michel Strogof</bibl>
<author>Beaumont and Fletcher</author>
<docAuthor>E. M. Forster</docAuthor>
```

2. Beispiel mit TEI-Elementen für die Auszeichnung von Autoren durch `<bibl>` und `<author>`:

[41] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-bibl.html` (besucht am 23.06.2016).
[42] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-author.html` (besucht am 23.06.2016).
[43] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-docAuthor.html` (besucht am 23.06.2016).
[44] Diese Beispiele sind aus den jeweiligen Dokumentationen der Elemente der TEI-Guidelines genommen. Die Online-Referenz wird pro Element angegeben.

```
<bibl>
    <author>
        <name>Taylor, Jane</name>
    </author>
</bibl>
```

3. Beispiel mit TEI-Elementen für die Auszeichnung von Autoren durch `<author>` und `<name>`:

```
<author>
    <name>
        <surname>Beaupaire</surname>
        <forename>Edmond</forename>
    </name>
</author>
```

4. Beispiel mit TEI-Elementen für die Auszeichnung von Autoren durch `<docAuthor>`:

```
<docAuthor>
    <forename>Perse</forename>
</docAuthor>
```

5. Beispiel mit TEI-Elementen für die Auszeichnung von Autoren mit Attributen:

```
<titleStmt>
    <title>The Rime of the Ancient Mariner:
        an annotated edition</title>
    <author xml:id="STC">Samuel Taylor Coleridge</author>
    <editor xml:id="JLL">John Livingston Lowes</editor>
</titleStmt>
```

Die Beispiele 1 bis 5 zeigen, dass unterschiedliche Informationen annotiert werden: Nachnamen, Initialen, Vornamen oder mehrere Autoren gleichzeitig. Um dies zu strukturieren, können weitere Elemente hinzugezogen und rekursiv angewendet werden. Die Spezifikation der Annotation erfolgt in Beispiel 2 durch die Elemente `<author>` und `<name>`. Das Element `<author>` wiederum kann nicht nur durch das Element `<name>`, sondern auch durch die Elemente `<surname>` und `<forename>` spezifiziert werden (vgl. Beispiel 3). Wie Beispiel 4 zeigt, ist die Anwesenheit des Elementes `<forename>` keine Voraussetzung dafür, dass das Element `<surname>` mit annotiert wird. Alle diese Vorgehensweisen fallen direkt unter die Richtlinien der TEI. Die verschiedenen Annotationsweisen sind unter dem Begriff der Personenannotation beschreibbar, hinzukommen im Falle der TEI noch mögliche Attribute

der Elemente. Zusätzlich können Elemente wie `<docAuthor>` auch mit allgemeineren Elementen `author` annotiert werden und nur die gesamte Struktur der Annotation zeigt auf, dass es sich bei dem annotierten Autoren um den Verfasser eines Dokumentes, der auf einem Titelblatt genannt wird, handelt.[45]

Eine Korpusdokumentation, die schlicht angibt, dass Autoren nach den TEI-Guidelines ausgewiesen werden, ist nicht aussagekräftig genug. So wird einerseits die getroffene Auswahl aus einer der verschiedenen Kategorisierungen in TEI nicht dokumentiert (z. B. Autor mit oder ohne Vor- und Nachnamen). Wenn andererseits Korpora in eine Anwendung wie einem Suchsystem eingelesen werden sollen, ist es relevant zu wissen, welche Elemente in welcher Extensible Markup Language (XML)-Struktur genutzt werden.

Die Annotationsbeispiele der TEI und der Wortartenannotationen mit dem STTS zeigen, dass verschiedene Auslegungen und Umsetzungen eines Standards existieren, unabhängig davon, ob automatisch, semi-automatisch oder manuell annotiert wird. Selbst die Ausweisung von *Autoren* kann also ähnlich wie bei *Wortarten* in Abhängigkeit von dem Korpus oder der Forschungsfrage unterschiedliche Kategorisierungen und Annotationen fordern.

Hier ist entscheidend, dass Annotationen, unabhängig von dem, was sie beschreiben, in dieser Arbeit als `Interpretationen` verstanden werden und damit in einem engeren Fachbezug und auch in einem engen Bezug zum Korpus und zur Forschungsfrage stehen können. Es werden viele verschiedene Schemata entwickelt und in Korpora z. T. unterschiedlich als empirische Basis für die Forschung genutzt. Dabei gibt es häufig verwandte oder ähnliche interpretatorische Kategorien, die sich dennoch in ihren Ausprägungen oder Ausweisungen stark unterscheiden können und somit schwer über deren theoretischen Konzepte, wie *Wortart* oder *Autor*, zusammenfassen und beschreiben lassen. Für die Dokumentation eines Korpus heißt dies, dass eine genauere Aufschlüsselung der Annotationsart, -werte und -weise wichtige, distinktive Eigenschaften von Korpora sind und diese in Betracht gezogen werden müssen. Dies muss in einer Korpusdokumentation berücksichtigt werden (Kapitel 6).

2.4 Korpusarchitektur

Mit einer Korpusarchitektur wird beschrieben, auf welche Art und Weise die Annotationen in dem digitalisierten sprachlichen Material ausgewiesen werden und wie sie

[45] Zu einer genaueren und umfangreicheren Diskussion der Anwendung der TEI Guidelines vgl. Kapitel 7.

sich auf das sprachliche Material und auf sich selbst beziehen können. Die verschiedenen Korpusarchitekturen ergeben sich aus der `Tokenisierung`, den `Annotationskonzepten`, den `Formaten` und den `Metadaten`. In diesem Abschnitt möchte ich nur auf ein paar grundlegende Ansätze eingehen, ohne eine umfassende Diskussion und Evaluation durchführen zu wollen. Herausgearbeitet werden sollen Konzepte und Eigenschaften von Korpora, die für deren Dokumentation relevant sind.

2.4.1 Tokenisierung

Die Zerlegung des sprachlichen Materials in Einheiten, die `Tokenisierung`, spielt eine zentrale Rolle bei der Korpusarchitektur. `Tokenisierung` meint die Zerlegung des digitalen Sprachmaterials in kleinste technische Einheiten, den `Tokens`. Viele Konzepte können hinter einem Token stehen. Ein Token kann z. B. für ein Zeichen, eine Silbe, ein Wort, ein Satz oder Absatz umfassen (vgl. Kuebler und Zinsmeister 2015: 7). Häufig werden graphematische Wörter und Satzzeichen als Token eines Korpus verstanden (vgl. Schmid 2008). Bei Sprachen ohne graphematische Wörter oder bei Transkripten gesprochener Sprache, die sprachliche Signale wie `Disfluencies` (Lickley 2015) enthalten, die nicht als klassische graphematische Wörter interpretiert werden können, ist die Zerlegung auf weitere, andere Kriterien angewiesen. Mit einer festen, nach bestimmten Kriterien durchgeführten Tokenisierung besteht beispielsweise die Möglichkeit, klare Vorhersagen über die Anzahl, Reihenfolge und die jeweils getrennt oder zusammen annotierten Einheiten in einem Korpus zu machen. Es gibt auch Korpusarchitekturen, die mehrere konzeptuelle Tokenisierungen zulassen, die dann als `multiple Segmentierungen` bezeichnet werden (Krause und Zeldes 2016; Krause et al. 2012).

Ob also eine Tokenisierung oder multiple Segmentierung im Korpus vorgenommen wurde, ist demnach für die Annotation relevant. Für historische Korpora ist dies besonders wichtig, wie Abschnitt 2.7 genauer zeigen wird. Nicht alle Korpora besitzen eine solche feste Tokenisierung, häufig fehlt diese beispielsweise bei Editionskorpora. Die Art der Tokenisierung hat einen Einfluss auf die Annotationskonzepte (vgl. Abschnitt 2.4.2), die in einem Korpus angewandt werden können, wie folgendes Beispiel in Tabelle 2.2 illustriert.

Tabelle 2.2: *Verschiedene Tokenisierungen für den Satz „Das gibt's nicht mehr.“*

tok1	Das	gibt's			nicht	mehr.		(4 Einheiten)
tok2	Das	gibt's			nicht	mehr	.	(5 Einheiten)
tok3	Das	gibt'		s	nicht	mehr	.	(6 Einheiten)
tok4	Das	gibt		's	nicht	mehr	.	(6 Einheiten)
tok5	Das	gibt		es	nicht	mehr	.	(6 Einheiten)
tok6	Das	gibt	'	s	nicht	mehr	.	(7 Einheiten)

Die Tabelle 2.2 zeigt, wie der Satz *Das gibt's nicht mehr.* nach unterschiedlichen Kriterien tokenisiert werden kann. Je nachdem, wie tokenisiert wird, ändert sich die Art und die Anzahl der Einheiten. Tok1 berücksichtigt allein die Leerzeichen zwischen den anderen Zeichen, tok2 trennt zusätzlich Satzinterpunktionszeichen. In tok3 wird das reduzierte Morphem *s* abgetrennt, welches in tok5 als *es* realisiert wird. Tok6 trennt Interpunktionszeichen *.* (Punkt) und *'* (Apostroph) sowie das reduzierte Morphem *s* ab. So kann in tok1 nicht direkt über die Tokens auf das Morphem *s* oder den Apostroph zugriffen werden[46], annotiert werden kann nur die zweite Einheit *gibt's* als Ganzes. Die Tokenisierungen tok3-tok6 erlauben dies hingegen auf verschiedene Weisen. Dadurch, dass die verschiedenen Tokenisierungen unterschiedlich viele Einheiten produzieren, variiert die Menge der Einheiten, die häufig als Normalisierungsgröße genutzt wird, und die Reihenfolge der Einheiten. Letzteres ist beispielsweise für die Untersuchung von Wortreihenfolgen interessant. So steht in tok1 an dritter Stelle *nicht*, in tok4 *'s* und in tok5 *es*, in tok6 der Apostroph. Diese Tokenisierungen verstehe ich hier als *feste* Tokenisierung: die Art der Einheiten ist festgelegt. Eine flexible Tokenisierung legt die Art und Anzahl der Einheiten nicht selbst konkret fest. Ein Beispiel dafür können TEI-Korpora sein, die dann Annotationen auf ad hoc festgelegten Einheiten wie in Abschnitt 2.3.2 erhält.

Wie tokenisiert wird, kann von Korpus zu Korpus unterschiedlich sein. Die Tokenisierung hat dann einen Einfluss, welche Annotationskonzepte im Korpus realisiert werden können. Das müssen Forscherinnen und Forscher wissen, wenn sie ein Korpus nutzen möchten. Diese Eigenschaft muss damit in einer Korpusdokumentation zum Zweck der Wiederverwendung berücksichtigt werden.

2.4.2 Annotationskonzepte

Für die Ausweisung von Kategorien wie in Abschnitt 2.3.2 und in Abschnitt 2.3.1 sowie für deren spätere Analyse werden unterschiedliche Annotationskonzepte ge-

[46] Bei genauer Kenntnis der Daten kann hier die Verwendung von Mustersuchen mittels regulärer Ausdrücke die Kombination aus Apostroph und *s* finden.

nutzt. `Annotationskonzepte` beschreiben die Art und Weise, wie Kategorien in einem Korpus ausgezeichnet werden. Wenn ein Korpus eine feste Tokenisierung besitzt, dann kann beispielsweise in Form einer `Tokenannotation` jedem Token genau ein Wert, z. B. eine Wortartkategorie nach einem Schema wie dem STTS, zugewiesen werden. Wenn Werte gleich mehreren Tokens zugeordnet werden, spricht man häufig von `Spannenannotation`. Annotationskonzepte wie Spannen sind ebenfalls etabliert und in unterschiedlichen Korpustypen umgesetzt. Einen gute Einführung über Tokens und Spannen geben Bird und Liberman (2000) am Beispiel von Sprachkorpora.

Neben den flachen Spannen- und Tokenannotationen werden auch hierarchisch organisierte Konzepte wie `Bäume` beispielsweise für die Annotation von Syntax verwendet. Zwei Tokens können beispielsweise über zwei Kanten mit einen Knoten verbunden sein und erhalten einen gemeinsamen Wert über diesen Knoten. Kanten können ebenfalls einen Wert erhalten. Für diskontinuierliche Annotationen, die beispielsweise Referenzketten abbilden, werden häufig gerichtete `Pointer` verwendet. Pointer verweisen von einem Token auf ein anderes Token, ohne dass diese direkt aufeinander folgen müssen. Alle Annotationskonzepte können mit ganz unterschiedlichen Kategorien und Interpretationen belegt werden. Die Umsetzung dieser Annotationskonzepte kann entweder direkt in dem digitalisierten Text als Inline-Annotation oder vom Text getrennt als Standoff-Annotation erfolgen.

Abbildung 2.1 zeigt ein Beispiel von verschiedenen Annotationskonzepten mit Tokenisierung. Diese Darstellung richtet sich nach dem Salt-Modell für Annotationen (Zipser und Romary 2010), wobei hier aus Gründen der Vereinfachung der Basistext nicht mit berücksichtigt wurde (vgl. dazu auch Krause et al. 2016).

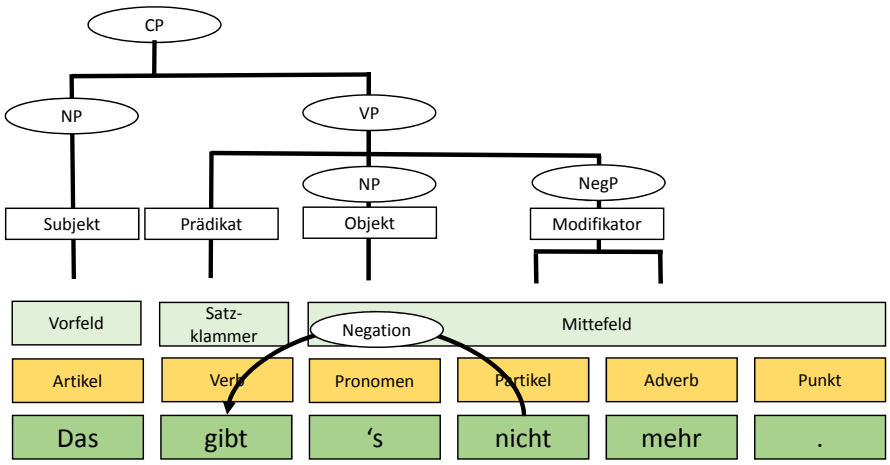

Abbildung 2.1: *Beispiel für verschiedene Annotationskonzepte. Der Satz Das gibt's nicht mehr. (tok4) erhält Token-, Spannen-, Baum- und Pointerannotationen.*

Abbildung 2.1 zeigt mit dem Beispiel tok4 aus Tabelle 2.2, wie Kategorisierungen in verschiedenen Annotationskonzepten umgesetzt werden können und welchen Einfluss die Tokenisierung auf die Annotation haben kann. Die Kategorien für Wortarten wie Artikel, Verb und Pronomen sind jedem Token zugewiesen (Tokenannotation). Durch die Trennung des Morphems *'s* vom *gibt* in tok4 ist hier eine getrennte Zuweisung von Wortartenkategorien möglich (Pronomen und Verb). In Form von Spannenannotationen werden einem oder mehren Tokens die Position im topologischen Feld zugewiesen. Die Tokens *'s*, *nicht*, *mehr* und *Punkt* werden dem *Mittelfeld* durch eine Spannenannotation zugeordnet. Die Struktur der Phrasen im Satz können beispielsweise durch eine hierarchische Annotation aus Knoten (Phrasen) und Kanten (Funktionen der Phrasen im Satz) abgebildet werden. So werden die NP (*'s*) und die NegP (*nicht mehr*) von einer VP dominiert. Jede dieser Phrasen erhält durch eine Annotation der Kanten Satzfunktionen wie *Objekt* oder *Modifikator*. Beziehungen, die diskontinuierlich aufgebaut sind, können mit Pointern annotiert werden. In diesem Beispiel wird die Negation von *gibt* durch *nicht* mit einem gerichteten Pointer angezeigt. Die Annotationskonzepte können so verschiedene Kategorisierungen abbilden. Das hier angeführte Beispiel zeigt nur eine Möglichkeit, wie Annotationskategorien und -konzepte zusammenspielen können. Denkbar sind beispielsweise auch Wortartartenkategorien, die nicht durch Tokenannotationen, sondern durch Pointer oder hierarchische Strukturen abgebildet sind.

Genau so können andere, nicht-linguistische Kategorisierungen mit diesen Annotationskonzepten umgesetzt werden. So können z. B. Texteinheiten, die sich auf einer Zeile oder Seite befinden oder wie in Abschnitt 2.3.2 Eigennamen darstellen, mit einer Spannenannotation auswiesen werden.

Es gibt eine Vielzahl an Annotationskonzepten, die auf unterschiedliche Weise technisch umgesetzt und eingesetzt werden (vgl. Bański und Przepiórkowski 2009; Gerdes 2013; Ide und Sudermann 2014; Zipser und Romary 2010). Daher existieren diverse Formate, die eine bestimmte Kombination aus Annotationskonzepten und Kategorien realisieren. Für die Wiederverwendung von Korpora ist es wesentlich, dass eine Korpusdokumentation Informationen über die Realisierung von Annotationskonzepten und -kategorien in einem oder mehren Formaten enthält.

2.4.3 Formate

Die verschiedenen Korpusarchitekturen, Kategorisierungen und Annotationskonzepte können in unterschiedlichen Formaten je nach Zweck, Annotation und Analyse umgesetzt werden. In der Linguistik gibt es Formate, die in aller Regel für ein bestimmtes Annotationskonzept verwendet werden, z. B. TIGER-XML für Tokens und Bäume (Romary et al. 2015), CoNLL für Tokens und Pointer (Hajič, Jan and Ciaramita, Massimiliano and Johansson, Richard and Kawahara, Daisuke and Mart\'ı, Maria Ant\'onia and M\'arquez, Llu\'is and Meyers, Adam and Nivre, Joakim and Pad\'o, Sebastian and Štěp\'anek, Jan and Straň\'ak, Pavel and Surdeanu, Mihai and Xue, Nianwen and Zhang, Yi 2009; Nivre et al. 2004)[47] und ELAN-XML (Wittenburg et al. 2006) sowie Extensible Markup Language for Discourse Annotation (EXMARaLDA)-XML (Schmidt und Wörner 2009) für die Realisation von ausschließlich Token- und Spannenannotation. Alle genannten Beispiele sind Formate, die fest tokenisierte Korpora abbilden. Das Format der TEI ist TEI-XML, welches in flexibel tokenisierten Korpora Konzepte wie Spannen und Pointer abbildet.

Zipser (2014) zeigt, wie divers allein in der Linguistik das Formatspektrum ist. Beispielsweise gibt es eine Vielzahl an XML-basierten Formaten für die Erstellung von Korpora (Dipper 2005; Heid et al. 2010; Romary et al. 2015; Schmidt und Wörner 2009; TEI Consortium 2015) und an CSV-basierten Formaten (Krause und Zeldes 2016; Nivre et al. 2004). Ebenso werden auch proprietäre Formate (Hennig 2013b), JSON-basierte Formate (Vertan et al. 2016) und graphbasierte Lösungen (Ide und

[47] `http://ilk.uvt.nl/conll/#dataformat` (besucht am 09.06.2016).

Sudermann 2014) genutzt. Weiterhin können Korpora in mehreren Formaten vorliegen, vor allem wenn sie in einer Mehrebenenarchitektur mit unterschiedlichen Annotationskonzepten vorliegen (Lüdeling 2012; Odebrecht et al. 2017; Romary und Ide 2004). Weiterhin werden verschiedene Formate für unterschiedliche Abschnitte des Forschungsdatenzyklus wie Erstellen, Analysieren und Archivieren verwendet. So gibt es spezialisierte Formate, die die Annotation von Korpora unterstützen, und spezialisierte Formate, die die Analyse von Korpora ermöglichen. Um den Aufwand, der bei der Unterstützung möglichst vieler Formate entsteht, zu minimieren, bieten sich Konvertierungsframeworks wie Pepper (Zipser und Romary 2010) an. Um ein Korpus wiederverwenden zu können, müssen die Korpusdaten verarbeitet, das heißt technisch und menschlich ausgelesen werden können. Dies erfordert unter anderem eine Dokumentation der jeweiligen Korpora mit den verwendeten Annotationskonzepten in den genutzten Formaten.

Wortartenannotationen, wie sie in Abschnitt 2.3.1 vorgestellt werden, werden typischerweise als Tokenannotation verstanden (Abbildung 2.1). Das Annotationskonzept und die Kategorisierung können in beispielsweise TIGER-XML genauso wie in CoNLL und in EXMARaLDA-XML umgesetzt werden. Die Annotationsbeispiele der TEI (Abschnitt 2.3.2) zeigen, wie ein nicht fest tokenisierter Text in Form von öffnenden und schließenden Elementen das Konzept der Spannenannotation in dem XML-Format darstellt. Eine ähnliche Interpretation funktioniert auch für unäre Elemente der TEI wie `<lb>`[48], die nicht öffnen und schließen, wie das Beispiel 6 zeigt. Hier kann eine Spanne interpretiert werden, die sich zwischen den gleichen Elementen ergibt (vgl. Krause et al. 2013).

6. Beispiel mit unären TEI-Elementen für die Auszeichnung von Zeilenumbrüchen:

```
Das ist eine Zeile <lb/> das steht in einer Zeile <lb/>
weiterer Text in einer Zeile <lb/>
Noch eine Zeile <lb/> eine weitere Zeile ...
```

Neben der Spezialisierung auf ein bestimmtes Annotationskonzept oder auf eine bestimmte Funktion im Forschungsdatenzyklus sind Formate auch unterschiedlich stark regulierend oder frei: Einige Formate wie TIGER-XML (Romary et al. 2015) und Text Corpus Format (TCF) (Heid et al. 2010) legen die Anzahl und

[48] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-lb.html` (besucht am 30.06.2016).

die Bedeutung der Annotationskategorien fest, andere wie TEI-XML lassen Spielraum in der Serialisierung und wieder andere wie EXMARaLDA-XML (Schmidt und Wörner 2009) oder Potsdamer AUstauschformat Linguistischer Annotationen (PAULA)-XML (Dipper 2005) geben keinerlei solcher Beschränkungen vor.

Das Zusammenspiel aus Tokenisierung, Kategorisierungen in Form von Annotationskonzepten und Formaten erzeugt also unterschiedliche Korpusarchitekturen. Eine Mehrebenenarchitektur erlaubt potenziell unendlich viele Annotationen mit unterschiedlichen Konzepten. Dabei können die Annotationen je nach Format entweder technisch zusammenhängend oder getrennt zugewiesen werden. Eine Mehrebenenarchitektur mit Stand-off-Annotation, wie es beispielsweise Romary und Ide (2004) vorsehen, benötigt ein spezialisiertes Format wie PAULA-XML (Dipper 2005).[49]

Zusammengefasst lässt sich feststellen, dass verschiedene Annotationskategorien in verschiedenen Annotationskonzepten umgesetzt werden. Diese wiederum können in unterschiedlichen Formaten realisiert werden. So ist es relevant zu wissen, welche Annotationskonzepte und -kategorien in welchem Format oder in welchen Formaten eines Korpus abgebildet werden. Dies sind Eigenschaften von Korpora, die für deren Wiederverwendung eine wesentliche Rolle spielen.

2.4.4 Metadaten

Neben den Annotationen besitzen Korpora häufig `Metadaten`, die weitere Informationen beispielsweise zu den Äußerungensituationen oder zu den Sprecherinnen und Sprechern der natürlich-sprachlichen Äußerungen geben:

> Metadaten sind Daten, die verschiedene Aspekte einer Informationsressource beschreiben. Die Informationsressource kann z. B. ein Text sein, eine Textsammlung, eine Tonaufnahme oder ein Video. (Lemnitzer und Zinsmeister 2006: 44)

Für textbasierte Korpora können das allgemein Metadaten über Autor, Veröffentlichungsjahr und -ort der Texte sein, für Zeitungskorpora Angaben über die Ausgabennummern und die Zugehörigkeit zum Ressort ebenso wie über die Autoren der Artikel. Korpora der gesprochenen Sprache enthalten typischerweise Angaben

[49] Neben den Formaten müssen auch die Analysetools die jeweiligen Konzepte unterstützen, z. B. Krause und Zeldes (2016) und Herzog et al. (2015). Auf Annotations- und Analysesoftware wird die Arbeit kurz in Abschnitt 2.6 weiter eingehen.

über die SprecherInnen in Hinblick auf Alter, Geschlecht, Muttersprache oder Wohnort. Die Metadaten variieren also je nach Korpustyp (vgl. Abschnitt 4.2). Je nach Datengrundlage und ausgewähltem Format können die Metadaten direkt im Annotationsformat, in der Analyseumgebung hinterlegt werden oder werden separat in Fließtextdokumentationen oder eigenen Metadatenprofilen hinterlegt (Abney und Bird 2011; Lehmberg und Wörner 2008; Zinsmeister et al. 2008). Auf Metadaten allgemein wird in Kapitel 4 noch genauer und umfangreicher eingegangen, einzelne Ansätze zur Metadatenmodellen werden in Kapitel 5 diskutiert.

2.4.5 Korpusgröße

Korpora unterscheiden sich ganz abstrakt hinsichtlich zwei Dimensionen:

1. Größe: Mit Größe wird der Umfang der im Korpus enthaltenen Texte beschrieben.

2. Tiefe: Mit Tiefe wird der Umfang der im Korpus enthaltenen Annotationen beschrieben.

Dabei gibt es verschiedene Einheiten, die die Größe eines Korpus bemessen: die geläufigsten sind Anzahl der im Korpus befindlichen Tokens oder der Wortformen. Die Tiefe eines Korpus wird in dieser Arbeit mit der Anzahl an Annotationsvariablen und -arten bemessen – auch Annotationsebenen und -layers genannt. Beide Dimensionen sind wiederum von der Korpusarchitektur inklusive der Tokenisierung abhängig.

2.5 Forschungsprozess und Korpusarchitektur

Hervorzuheben ist, dass in dieser Arbeit nicht der Versuch unternommen wird, eine bestimmte Korpusarchitektur oder bestimmte Annotationskonzepte sowie Interpretationen grundsätzlich zu empfehlen. Nach der Argumentation von Lüdeling (2011) ist die korpuslinguistische Aufbereitung von Texten immer eine interpretative Methode und daher ist die Entscheidung für eine bestimmte Umsetzung immer abhängig von der Forschungsfrage.

Dieser Abschnitt hat gezeigt, dass Korpora unterschiedliche Architekturen, Formate und Annotationskonzepte und -kategorien nutzen. Je Korpus kann sich das Zusammenspiel der einzelnen Komponenten einer Korpusarchitektur unterscheiden, daher ist die Forschungsdatenlandschaft sehr heterogen.

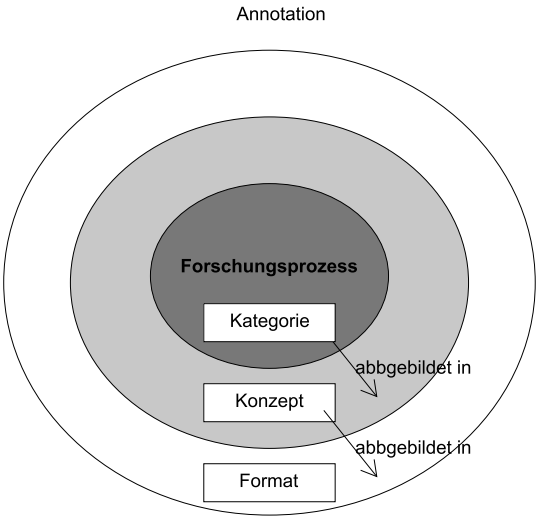

Abbildung 2.2: *Zusammenspiel von Annotationskategorien, -konzepten und -formaten bezüglich des Forschungsprozesses und der Korpusdokumentation. Die Grau-Weiß-Stufung zeigt die Nähe zum Forschungsprozess an. Je dunkler, desto zentraler ist die Komponente für den Forschungsprozess.*

Abbildung 2.2 zeigt, wie dieses Zusammenspiel in der Korpusdokumentation berücksichtigt werden wird. Die Annotationskategorien sind eine zentrale Komponente des Forschungsprozesses, die je Korpus sehr spezifisch dokumentiert werden muss. Die Annotationskonzepte, die diese Kategorien abbilden können, sind ebenfalls Teil des Forschungsprozesses, sie können aber mehrere unterschiedliche Kategorien abbilden und sind deshalb weniger eng mit einem Forschungsprozess verbunden. Die Formate, die diese Konzepte abbilden können, sind noch weniger korpus- oder forschungsspezifisch und daher weniger eng mit einem Forschungsprozess verknüpft. Ihre Dokumentation kann und wird zum einem großen Teil unabhängig von einzelnen Korpora erfolgen.[50] Die verwendeten Tools, die Formate auslesen können, stehen dann in einer wiederum weniger engen Verbindung zum konkreten Forschungsprozess eines Korpus und besitzen typischerweise ebenfalls eine korpusunabhängige Dokumentation. Daraus folgt, dass je enger eine Komponente mit dem Forschungsprozess verbunden ist, desto mehr bzw. umfangreicher muss sie in der Korpusdokumentation berücksichtigt werden.

Wie bestimmte Annotationen in einem Format abgebildet und in einer Anwen-

[50] Die wenigen Fälle, in denen ein Format für genau ein Korpus entworfen sind, bilden dazu eine Ausnahme.

dung bearbeitet werden können, ist eine interdisziplinäre Aufgabe, die durch die Einbindung in den Forschungsprozess einen hohen innovativen Charakter besitzen kann. In der Korpuslinguistik ist interdisziplinäre Arbeit integraler Bestandteil und InformatikerInnen, ComputerlinguistInnen und KorpuslinguistInnen arbeiten eng zusammen (Kytö 2011: 435). Auch andere geisteswissenschaftliche Fächer, wie sie Initiativen der Digital Humanities[51] unterstützen, arbeiten in gleicher Form.

Ein Korpus wird als ein Produkt des Forschungsdatenlebenszyklus verstanden. Tokenisierung, Annotationskategorien, Annotationskonzepte, Größe und Formate werden in verschiedenen Bearbeitungsschritten genutzt und stellen damit zu beschreibende Komponenten für diesen Forschungsdatenzyklus dar. Historische Korpora grenzen sich noch einmal von anderen Textkorpora durch die Art des digitalisierten sprachlichen Materials und dessen Anforderungen an eine Aufbereitung ab. Historische Texte stellen besondere Anforderungen in Bezug auf ihre Tokenisierung, Annotationskategorien (Transkription, Normalisierung) und Dokumentation (vgl. Abschnitt 2.7). Daneben ist der Forschungsdatenzyklus auch durch die Verwendung von Tools gekennzeichnet, die ein Format mit bestimmten Annotationskategorien und -konzepten für eine Bearbeitung oder Analyse auslesen können. Abschnitt 2.6 versucht eine kurze Zusammenfassung dieser Tools, um die für eine Korpusdokumentation, wie sie in dieser Arbeit vorgeschlagen wird, relevanten Informationen zum Lebenszyklus der Forschungsdaten zu identifizieren.[52] Diese Informationen sind für die in Kapitel 3 vorgestellten Wiederverwendungsszenarien relevant und werden in der Modellierung (Kapitel 6) berücksichtigt.

2.6 Korpusdatenverarbeitung

Analyseumgebungen, Konverterframeworks, Annotationstools und -pipelines wirken zu den verschiedenen Stadien des Korpus in seinem Lebenszyklus mit. Welche Anwendungen konkret und in welcher Kombination für das vorliegendes Korpus angewandt werden, ist daher ein wesentlicher Aspekt, da nur so der jeweilige Lebenszyklus nachvollzogen werden kann. Dieser Abschnitt stellt nur knapp Software zur Erstellung von Korpora wie Annotationstools und Verarbeitungspipelines vor und verweist auf die jeweiligen Dokumentationen, da der Schwerpunkt dieser Arbeit auf

[51] Wie European Association for the Digital Humanities (EADH), `http://eadh.org/`, The Alliance of Digital Humanities Organizations (ADHO), `http://adho.org/` oder Digital Humanities im deutschsprachigem Raum (DHd),`https://dig-hum.de/` (besucht am 04.08.2016).

[52] Diese Zusammenfassung erhebt keinen Anspruch auf Vollständigkeit.

der Dokumentation von Korpora und nicht auf der Dokumentation von Software liegt.

Um Korpora manuell zu erstellen, zu annotieren oder zu korrigieren, werden typischerweise Stand-alone-Tools (z. B. Druskat et al. 2014; Schmidt und Wörner 2009) oder webbasierte Annotationstools (z. B. Gerdes 2013; Muhie Yimam et al. 2013) genutzt.[53] Annotationstools ermöglichen AnwenderInnen die computergestützte Annotation der natürlichsprachlichen Äußerungen. Diese Anwendungen nutzen häufig ein eigenes Format und sind auf bestimmte Annotationskonzepte oder Korpustypen spezialisiert.[54] Wenn Korpora eine komplexe Architektur und damit mehrere Annotationskonzepte besitzen, werden für die Erstellung oder Annotation auch mehrere Tools genutzt (vgl. hierfür auch Abschnitt 2.4.3).

Neben der manuellen Annotation von Korpora werden im Rahmen der Natural Language Processing (NLP)-Forschung die maschinelle Ver- und Bearbeitung von natürlichsprachlichen Ressourcen jeder Art erforscht. Darunter fällt unter anderem die Entwicklung von Tools, die automatisch Sprache taggen oder parsen; also automatisch Annotationskategorien in einem Korpus ausweisen. Dazu können diese Tools die Zuweisung der gewählten Annotationsschemata und Annotationskonzepte für eine Sprache oder einen Ressourcentyp erlernen.[55] Häufig werden solche Anwendungen auf bestimmten Korpustypen (z. B. auf modernen Zeitungstexten) trainiert oder spezialisiert. Daneben existieren auch für historische Korpora spezialisierte Tools, die beispielsweise die manuelle, semi-automatische oder automatische Normalisierung unterstützen (Baron und Rayson 2008; Bollmann et al. 2012; Jurish 2010).[56]

Tagger sind typischerweise auf die Tokenannotationen von Kategorien für Wortarten oder Lemmatisierungen spezialisiert (Schmid 2008; Schmid und Laws 2008). Parser weisen syntaktische Informationen mit Pointer-Annotationen oder hierarchi-

[53] Stand-alone-Tools sind Anwendungen, die eigenständig auf einen lokalen PC als Desktopanwendungen installiert und genutzt werden können. Webbasierte Anwendungen laufen über einen Server, den Anwenderinnen und Anwendern selbst werden nur auf einem lokalen Client die Anzeige oder Ausgabe der Datenverarbeitung angezeigt.

[54] Vgl. z. B. Dipper et al. (2004) für eine Evaluation von Annotationstools für linguistische Annotation. Nicht immer können nur bestimmte Tools bestimmte Formate auslesen. Einige Korpusformate wie TEI-XML können auch mit generischen XML-Editoren wie Oxygen bearbeitet werden, vgl. `https://www.oxygenxml.com/` (besucht am 29.12.2016).

[55] Für eine erste Einführung und einen Überblick zu maschinellem Lernen in der Computerlinguistik vgl. Carstensen et al. (2010) und Jurafsky und J. H. Martin (2009). Solche Tools können auch regelbasiert arbeiten, womit der Schritt des Erlernens entfällt.

[56] Einen ersten Überblick zu NLP für historische Korpora gibt Piotrowski (2012). Abschnitt 2.7 geht noch einmal genauer auf die Besonderheiten bei der Annotation von historischen Korpora ein.

schen Baumannotationen aus (Chen und Manning 2014; Nivre et al. 2004). Solche Tools werden auch in Verarbeitungspipelines manchmal in einer virtuellen Umgebung zusammengefasst (Ferrucci und Lally 2004; Ide et al. 2016; Kok et al. 2015).

Die verschiedenen Verfahren zur Annotation von Korpora können kombiniert werden. Beispielsweise können die in Abschnitt 2.3.1 vorgestellten Wortartenannotationen nach dem STTS anstatt einer manuellen Zuweisung auch mit einem Tagger automatisch zugewiesen und danach manuell oder semi-automatisch in einem weiteren Tool (wie dem EXMARaLDA-Partitur-Editor) zusammen mit anderen Annotationen weiter bearbeitet werden, wie es z. B. im Fürstinnenkorrespondenzkorpus[57] umgesetzt wurde.

Wenn für die Umsetzung mehrerer Annotationskonzepte auch mehrere Tools und damit mehrere Formate für die Erstellung eines Korpus benötigt werden, dann muss ein Format zur Weiterverarbeitung in ein anderes überführt werden, wofür Konvertierungstools (z. B. Zipser und Romary 2010) benötigt werden.

Neben der Erstellung von Korpora ist deren Durchsuchbarkeit und Analyse ebenso ein Teil des Forschungsdatenzyklus und wird durch verschiedene Anwendungen unterstützt. Die Auswertung von Korpora kann mit allgemeineren Such- und Visualisierungstools wie Corpus Search, Management and Analysis System (COSMAS) (Institut für deutsche Sprache Mannheim 2007), Search and Visualization in Multilayer Linguistic Corpora (ANNIS) (Krause und Zeldes 2016), Corpus Query Processor (CQP)[58] oder mit spezielleren Anwendungen beispielsweise Text-Retrieval-Tools für Ägyptologen (Iglesias-Franjo und Vilares 2016) sowie mit Anwendungen für Statistik wie R (R Core Team 2016)[59] erfolgen. Analysewerkzeuge besitzen häufig eigene Formate, so dass das Korpus aus einem Annotationsformat in ein Format des Analysewerkzeugs konvertiert werden muss. Der gesamte Teil der Analyse und Auswertung eines Korpus ist also auch ein Teil des Lebenszyklus von Forschungsdaten. Eine Voraussetzung für die Suche nach Annotationen in Korpora sind wiederum Korpusdokumentationen. So werden die Analysemöglichkeiten wie die Bearbeitungsmöglichkeiten von Korpora bei der Korpusdokumentation und der Erarbeitung der Wiederverwendungsszenarien berücksichtigt.

Zusätzlich zur Erstellung und Analyse von Korpora ist auch deren Publikation ein wesentlicher Bestandteil des Forschungsdatenzyklus. So speichern Repositorien und Archive wie das Long-term Access and Usage of Deeply Annota-

[57] Vgl. die Korpusdokumentation `http://hdl.handle.net/11022/0000-0000-82A0-7` (besucht am 23.11.2016).
[58] `http://cwb.sourceforge.net/files/CQP_Tutorial/` (besucht am 08.10.2016).
[59] `https://www.r-project.org/` (besucht am 08.10.2016).

ted Information (LAUDATIO)-Repositorium (Krause et al. 2014), das Archiv Textgrid (Neuroth et al. 2011), die Infrastruktur Collaborative European Digital Archive Infrastructure (CENDARI) und das Hamburger Zentrum für Sprachkorpora (HZSK) die unterschiedlichsten Korpora und stellen zusätzlich Metadaten zu den Korpora zur Verfügung. Auf die verschiedenen Ansätze zur Korpusdokumentation geht Kapitel 5 ein.

So besitzen Korpora analog zu ihren unterschiedlichen Annotationskonzepten und Annotationsarchitekturen auch ganz unterschiedliche Verarbeitungswege. Welche Annotation dann wie und mit welchem Tool und in welchem Format ausgewiesen, analysiert oder archiviert wird, sind damit wesentliche Informationen für die Korpusdokumentation.

In diesem Anschnitt wurden nur einige der Hilfsmittel und Werkzeuge stellvertretend genannt, die an unterschiedlichen Punkten des Lebenszyklus von Forschungsdaten (Erstellung, Annotation, Speicherung, Analyse) ansetzen. Die jeweilige Bearbeitungsschritte eines Korpus können mit ganz unterschiedlichen Tools durchgeführt werden. Entscheidend ist, dass die Hilfsmittel, die dazu notwendig sind, in der Korpusdokumentation berücksichtigt werden.

2.7 Historische Korpora

Einen besonderen Typ Korpus stellen historische Korpora dar. Sie zeichnen sich besonders durch die Anforderungen aus, die historische Texte an die Datenerstellung und -analyse stellen.
Ein historisches Korpus wird hier als ein Subtyp textbasierter Korpora verstanden. Historische Korpus werden erstellt, um eine vergangene Sprachstufe zu repräsentieren und zu untersuchen oder um Sprachwandel zu erforschen (Claridge 2008: 242).[60] Übertragen auf andere Fachbereiche, die mit historischen Korpora arbeiten, sind der Untersuchungsgegenstand die vergangenen Phasen einer oder mehrerer literarischen, politischen oder soziologischen Epochen. Viele (kritische) digitale Editionen, wie sie beispielsweise die Literaturwissenschaft, die Philologie oder auch die Geschichtswissenschaft erzeugen, sind Korpora in dem oben genannten Sinn (für eine Einführung in digitale Editionen vgl. Apollon et al. 2014). In dieser Arbeit sind Editionen ebenfalls Korpora und damit als Sammlungen natürlicher geschriebener Sprache, die digital vorliegen und mit Annotationen versehen sind. Die Annotatio-

[60] Für einen Überblick über historische Korpora vgl. z. B. Kytö (2011) und Gippert und Gehrke (2015).

nen in diesen digitalen Editionen enthalten dann typischerweise Kategorien, die die graphischen und editorialen Eigenschaften des Dokumentes ausweisen (Cover und Robinson 1995). Im Folgenden wird einheitlich der Begriff `Korpus` auch für `digitale Editionen` verwendet.

Die historischen Texte, die in diesen Korpora verarbeitet werden, stellen spezielle Anforderungen an die Korpusaufbereitung und -dokumentation. Die Einordnung, was ein `Text` ist, wie seine Überlieferungsgeschichte ist und auf welche konkreten Quellen zurückgegriffen werden kann, sind wesentliche Fragen. Darin spiegeln sich auch die verschiedenen Auffassungen, was ein Primärtext ist (Abschnitt 2.7.1), genauso wie die verschiedenen Ansätze zur Digitalisierung und Annotation von historischen Texten (Abschnitt 2.7.2). Daraus ergeben sich konzeptionelle und technische Anforderungen, die wiederum in der Korpusdokumentation besonders berücksichtigt werden müssen.

2.7.1 Historische Texte in Korpora

Ein wesentliches Merkmal des Korpustyps `historisches Korpus` ist die Fokussierung auf das Dokument, von Pitti (2004) *dokument-zentrisch* genannt. Das historische sprachliche Material – der `Text` – ist der konzeptionelle Kern des Korpus, auf Grundlage dessen weitere Analysen auf Basis von Annotationen getätigt werden.

Die Definition von `Text` oder `Dokument` ist nicht trivial. Die Functional Requirements for Bibliographic Records (FRBR)[61] hat eine Vier-Level-Beschreibung für (digitale) Dokumente entwickelt. Diese Vier-Level-Beschreibung fokussiert sich auf eine abstraktere Sichtweise auf ein Forschungsdatum – ein Dokument, in dem die Beschreibungskomponenten aus den verschiedenen Leveln für eine bibliographische Dokumentation festgehalten werden. Damit kann dieser Ansatz als eine Art Repräsentation der bibliographischen Eigenschaften von Dokumenten verstanden werden (Zeng und Qin 2016: 164).

> The entities in the first group [...] represent the different aspects of user interests in the products of intellectual or artistic endeavour. The entities defined as work (a distinct intellectual or artistic creation) and expression (the intellectual or artistic realization of a work) reflect intellectual or artistic content. The entities defined as manifestation (the physical embodiment of an expression of a work) and item (a single exemplar of

[61] `http://www.ifla.org/VII/s13/frbr/` (besucht am 02.12.2016).

> a manifestation), on the other hand, reflect physical form. (IFLA 2009: 13)

Nach diesem Ansatz kann das intellektuelle *Werk* (wie eine Geschichte) dabei in mehreren *Expressionen* (Versionen dieser Geschichte in beispielsweise verschiedenen Sprachen) umgesetzt werden. Solche *Expressionen* wiederum können in mehreren *Manifestationen* (gedrucktes Buch) existieren und diese wiederum in mehreren *Exemplaren* (einzelne Bücher). Jeder dieser Levels bezieht sich damit auf etwas anderes und muss mit unterschiedlichen Metadaten im Rahmen einer Korpusdokumententation beschrieben werden.[62] Damit kann mit `Text` jeder dieser Level angesprochen werden. Mit Dokument kann eine Expression, ein Manifestation oder ein Exemplar gemeint werden.

Eine Dokumentation gemäß diesem Vier-Level-Modell kann dann gerade bei historischen Texten einen hohen Komplexitätsgrad aufweisen. Deren Texthistorie kann typischerweise weitaus komplexer sein als bei modernen schriftlichen Erzeugnissen. Das historische Werk liegt als gedruckte Fassung, als Manuskript oder als Edition vor. Nicht aus allen Manifestationen kann die Expression oder das Werk abgeleitet werden. Bei historischen Texten (Werk) kann es beispielsweise einen Abstand zwischen Werkproduktion oder Werkrealisation (Expression) und der Textüberlieferung (Manifestation) geben. Ein Werk aus dem 9. Jahrhundert könnte so in einer Edition aus dem 19. Jahrhundert vorliegen, wobei hier das Werk oder Teiles des Werks (z. B. seine originale Sprache) möglicherweise nicht mehr komplett rekonstruiert werden kann. Daher sind die Überlieferungsgeschichte, die aufzeigt, welche Texte wann rezipiert worden sind, und die Editionsphilologie, welche sich auf den Inhalt der Texte bzw. den Text selbst fokussiert, wichtig für den Forschungsprozess und für die Dokumentation von Texten (vgl. z. B. Fleischer und Schallert 2011; Hübner 2006; Weddige 2006). Hinzu kommen kann, dass Exemplare unterschiedlich gut erhalten oder vollständig vorliegen. Beispielsweise könnten Seiten eines Exemplars fehlen oder teilweise beschädigt oder beschmutzt sein, sodass eine Texterfassung zu unterschiedlichen Ergebnissen bei verschiedenen Exemplaren kommen kann. Eine Dokumentation müsste darüber ebenfalls Informationen beinhalten.

Ein weiterer zu berücksichtigender Aspekt von Texten sind die Eigenschaften der unterschiedlichen Manifestationen. Historische Korpora können sich auch auf eine traditionelle, analoge Publikationsformen und deren Segmente beziehen, wie beispielsweise eine Monographie oder innerhalb eines Textes Seiten, Zeilen, Kapitel

[62] Vgl. Romary (2013), Solodovnik (2011) und Zeng und Qin (2016) für eine detailliertere Beschreibung dieses Ansatzes.

oder Abschnitte. Nicht immer werden ganze Publikationseinheiten digital aufbereitet, häufig sind es auch Auszüge oder einzelne Kapitel und Textfragmente. Für die Beschreibung und Auswertung eines Korpus ist es wichtig zu wissen, wie das sprachliche Material ausgesucht und ggf. gekürzt und gesampelt wurde (vgl. z. B. für linguistische Korpora Lüdeling 2011). Das Korpusdesign von historischen Korpora stellt sich damit den gleichen Herausforderungen wie bei anderen Korpustypen (Abschnitt 2.2).

Welche Art der Publikation (Manifestation) herangezogen wird, hat auch einen Einfluss auf die Digitalisierung und Annotation des historischen Textes und damit auf die Transkription und ggf. auf die Normalisierung. Im Rahmen einer Korpuserstellung müssen daher folgende Herausforderungen erfasst und gelöst werden:

> [S]pelling variation presents a problem for automatic annotation and searching of historical texts, and there has been some tension between the respect felt by historical linguists for the source text and the demands set by searchability. [...] To normalise or not to normalise, that was the hotly debated question for quite some time, with those remaining in the minority who advocated the need for normalised versions of the text. (Kytö 2011: 439)

Eine Digitalisierung (z. B. manuelle oder automatische Transkription) einer modernen Edition eines historischen Textes birgt andere Aufgaben als die Digitalisierung eines historischen Manuskriptes. Eine Edition kann beispielsweise einen gut lesbaren, gedruckten Text enthalten, der aber durchaus bereits editorische Anpassungen wie Normalisierungen durchlaufen hat. Im Gegensatz dazu kann der historische Text in Manuskripten lückenhaft, beschädigt und damit nicht gut lesbar sein. Dessen Digitalisierung ist nicht ohne Weiteres möglich. Ein weiterer Aspekt, der durch die Wahl der historischen Vorlage beeinflusst ist, ist die meist hohe Schreibvarianz von historischen im Vergleich zu modernen Texten, die eine automatische Verarbeitung mit NLP-Tools und ein Durchsuchen des Textes selbst erschwert. Die Aufbereitung von historischen Texten muss also zwischen einer originalgetreuen Darstellung und einer Normalisierung entscheiden. Beide Varianten forcieren jeweils Entscheidungen anders, was genau aus dem Original (analogen Vorlage) übernommen werden kann. Wie genau die Korpora diese Frage lösen können, zeigt Abschnitt 2.7.2.

Darüber hinaus kann es von historischen Vorlagen verschiedene Varianten geben, die alle als Expressionen desselben Werks unterschiedliche Eigenschaften besitzen.

> For instance, early imprints of one and the same work may differ in details owing to compositors having made changes to the type in individual copies. (Kytö 2011: 431)

Diese können beispielsweise in Parallelkorpora zusammengestellt werden, in dem alle vorhandenen Texte parallel aufbereiten werden:

> The different texts are not just one-to-one copies from some source(s) but show considerable variation and, at least in parts, seem to be independent creations.[...] As a consequence, we treat all texts equally, in contrast to most other historical text editions. (Dipper und Schultz-Balluff 2013: 28)

Spätestens hier knüpft sich die Frage an, welcher Text der `Primärtext` oder das `Primärdatum` ist. Je nach Blickwinkel kann dann das Werk, eine Expression, eine Manifestation in Form von Editionen oder Manuskripten oder die korpuslinguistische Transkription als eine Art `Primärtext` oder `Primärdatum` verstanden werden. In enger Beziehung zu dem `Primärdatum` oder der `Primärquelle` steht dann die `Sekundärquelle`:

> For historians, historiography signals a shift from "primary" sources – often archival ones – to "secondary" sources – or the historical arguments, interpretations, and interventions that use the archives to mount claims about the past. Of course, this distinction is rather artificial: today's "secondary" sources often become tomorrow's "primary" ones; what seems in the archive to offer direct access to the past is itself fundamentally representational and interpretive in nature already. (Kramer 2014: Abschnitt zu *Historiography*)

Auf einer abstrakten Ebene, wie sie Kramer (2014) aufgreift, werden die verschiedenen Perspektiven, was eine Primärquelle und was ein Sekundärquelle ist, thematisiert. Die Definition, was jeweils unter Primärtext verstanden wird, ist also in vielen Fällen nicht klar und ist durch den Forschungsprozess bedingt (vgl. auch Claridge 2008; Himmelmann 2012). Im Rahmen der Korpustypen werden ebenfalls unterschiedliche Definitionen für primäre und in der Folge sekundäre Datengrundlagen diskutiert (vgl. Abschnitt 2.2).

Für historische Korpora besteht dann ein Zusammenhang zwischen der Frage nach dem Primärtext, der komplexen Überlieferungsgeschichte eines historischen Textes

und der Vier-Level-Beschreibung sowie der korpuslinguistischen Aufbereitung (Abbildung 2.3).

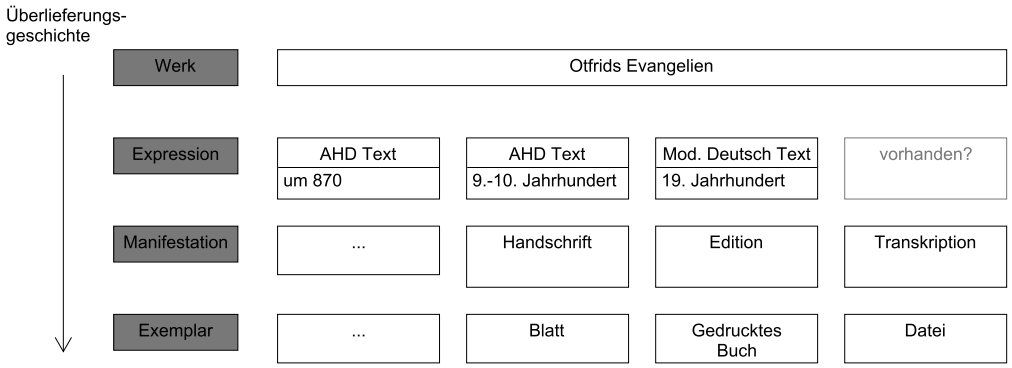

Abbildung 2.3: *Die Interaktion zwischen Werk, Expressionen, Manifestationen und Exemplaren sowie der Transkription eines historischen Texts.*

Abbildung 2.3 illustriert anhand eines kleinen Beispiels, wie eine solcher Zusammenhang interpretiert werden kann. Ein Werk wie *Otfrids Evangelienbuch* besitzt eine althochdeutsche Expression, deren Manifestation nicht bekannt ist. Daneben gibt es eine zweite althochdeutsche Expression, die in Form eines Manuskriptes manifestiert ist. Eine nächste Expression in modernem Deutsch liegt in Form einer Edition desselben Werks vor. Jede dieser Expressionen besitzt eigene Besonderheiten, die untereinander verglichen werden können und damit nicht als dasselbe interpretiert werden. Ob jede Edition eine eigene Expression besitzt, ist bei diplomatischen Editionen bereits fraglich. Ist der Text in einer solchen Edition bereits so eigenständig, dass er mit dem Text aus einer Handschrift verglichen werden kann? Die Überlieferungsgeschichte des Werks kann darüber hinaus je nach Datenlage beschreiben, welche der verschiedenen Realisationen voneinander abhängen.

Wenn nun noch das historische Korpus miteinbezogen werden soll, dann stellt sich die Frage, wo die Transkription eingeordnet werden kann. Ist eine diplomatische Transkription eine Expression des Werks selbst oder eine Manifestation einer bereits vorhandenen Expression? Eine diplomatische Transkription einer Edition kann als eine zweite Manifestation des modernen deutschen Textes oder als eigene Expression, die sich linguistisch von den anderen Expressionen unterscheidet, interpretiert werden. Transkribiert werden kann im Prinzip jede Expression, die eine Manifestation besitzt, die als physisch vorhandene Vorlage genutzt werden kann. Wenn die

Transkription sehr frei erfolgt und Modifikationen der Vorlage zulässt (bis hin zur Normalisierung oder Paraphrasierung), dann müsste sie eine eigene Expression besitzen. Die Frage ist, inwieweit eine Transkription als Manifestation ohne eigene Expression wirklich interpretierbar ist. Was ist für die Transkription primär? Ab wann ist eine Transkription nicht mehr „diplomatisch genug“ und schon eher normalisierend oder so unabhängig, dass eine eigene Expression angenommen werden muss?

Diese Arbeit unternimmt keinen Versuch, Primärquellen oder Primärtexte einheitlich zu definieren, sondern vertritt die Position, dass die Entscheidung, was ein Primärtext oder eine Primärquelle in einem Korpus ist, allein über die jeweilige Forschungsfrage und die Theorie zur Forschung motiviert wird (Odebrecht 2014). Daher wird hier eine andere, technisch-abstrakte Perspektive in der Modellierung für `Text` (Kapitel 6) eingenommen, die sich auf die Korpusarchitektur, die für historische Korpora einheitlich beschrieben werden soll, stützt.

2.7.2 Annotation historischer Korpora

Je nachdem, welche Art Expression oder Manifestation eines historischen Textes als Vorlage für die Digitalisierung genommen wird, werden verschiedene Entscheidungen für deren Richtlinien getroffen. Die Digitalisierung bzw. die Transkription hat einen großen Einfluss auf die gesamte Korpusarchitektur (Abschnitt 2.4). Ganz unterschiedliche Verfahren und Richtlinien werden beispielsweise an historischen Korpora des Deutschen, gerade in Bezug auf die Texterkennung/Transkription, Normalisierung und das Tagging angewendet, wie z. B. Barteld et al. (2016), Bartsch et al. (2011), Bollmann et al. (2012), Demske (2007), Donhauser (2015), Durrell et al. (2007), Hennig (2013b), Jurish (2010), Odebrecht et al. (2017), Petrova et al. (2009) und Solms und Wegera (1998). Genau wie in historischen Korpora des Deutschen ist eine ähnliche Vielfalt und Diversität bei Korpora anderer Sprachen wie dem Englischen oder bei Korpora romanischer Sprachen zu beobachten (vgl. Archer et al. 2015; Burr et al. 2015; Kroch und Taylor 2000; Kytö 1996).

Unter Digitalisierung verstehe ich hier die Transkription des Textes, deren Produkt dann unabhängig von der Vorlage und den Richtlinien bestimmte Eigenschaften besitzt (Pitti 2004: 477):

- Ein Text ist im Prinzip eine Folge von Zeichen.
- Es gibt uneinheitliche Definitionen für enthaltene Texteinheiten, wie Wörter.

- Diese Einheiten sind seriell.
- Ein Text kann semi-strukturiert sein, z. B. durch Markup.

Jede Digitalisierung (Transkription) erzeugt damit eine serielle Abfolge von Zeichen, die dann ganz unterschiedlich tokenisiert werden kann (vgl. Abschnitt 2.4.1). Wie bei anderen Annotationen können Transkription und Normalisierung auf verschiedenen Prinzipien basieren und werden dann in Annotationsrichtlinien formuliert. Damit sind sie wie andere Annotationen als Interpretationen zu verstehen, die aus einem Forschungsprozess heraus entwickelt werden. Bei Transkriptionen ist eine diplomatische Arbeitsweise typisch aber nicht notwendig. Eine diplomatischen Arbeitsweise legt ganz allgemein den Fokus darauf, den Text wird so nah wie möglich an der Vorlage (Edition, Handschrift etc.) zu transkribieren. Wie bei anderen Kategorisierungen kann dies ganz unterschiedlich gelöst werden. Dabei muss beispielsweise in den Richtlinien festgehalten werden, wie mit nicht lesbaren (beschädigten) Textabschnitten, Sonderzeichen, Schreibvarianten, der Markierung von Zeilenumbrüchen oder Textmarkup wie Fettdruck oder Farben umgegangen werden kann.

Die folgenden Beispiele aus dem Anselm-Korpus (Abbildung 2.4) (Dipper und Schultz-Balluff 2013) [63], Historisches Predigtenkorpus zum Nachfeld (HIPKON) (Coniglio et al. 2014, 2016) (Abbildung 2.5)[64] und RIDGES (Abbildung 2.6)[65] zeigen, wie unterschiedlich die diplomatische Transkription von Zeilenumbrüchen erfolgen kann.

[63] Belegreferenz: Sigle: Le, Aufbewahrungsort: Universiteit Leiden – Bibliotheken Signatur: Hs. Ltk. 226. `http://www.ruhr-uni-bochum.de/wegera/Le_Leid.Ms.Ltk226.pdf` (besucht am 02.01.2017).

[64] Coniglio, Marco; Donhauser, Karin; Schlachter, Eva; HIPKON (Version 1.0), Humboldt-Universität zu Berlin. SFB 632 Teilprojekt B4.. http://www.sfb632.uni-potsdam.de/. http://hdl.handle.net/11022/0000-0000-2D18-4. Belegreferenz: `https://korpling.org/annis3/?id=e93929e5-2b16-4b47-b449-a3de80939453` (besucht am 02.01.2017).

[65] Belegreferenz: `https://korpling.org/annis3/?id=76a4fa6e-b1ac-4d59-8ec2-37dbe459f114` (besucht am 02.01.2017).

Le1_84v,14 rouwe als doe ſi hem ſach ghe=
Le1_84v,15 ſele<...> beſpuwenen̄ hanghen an=
Le1_84v,16 den cruce en̄ ſijn herte ontwee
Le1_84v,17 ſteken ſinen aderen ſcoren ende
Le1_84v,18 ſinen oghen vergaen Dit heuet
Le1_84v,19 ſi al in horen herte beſloten hier(=)
Le1_84v,20 om heuet ſi ſo groten rouwe

Abbildung 2.4: *Ausschnitt aus dem Anselm-Korpus: Transkription von Zeilenumbrüchen, mit Erhaltung der Worttrennung und zusätzlicher Hinzufügung von Trennungszeichen.*

daz er ſin lîvte ge(=)heilet hat .

Abbildung 2.5: *Ausschnitt aus dem HIPKON-Korpus: Transkription von Zeilenumbrüchen, ohne Erhaltung der Worttrennung und mit Markierung durch Trennungszeichen.*

ziehe diſs kraut mit der wur
tzel aus / vnd leg das vber nacht ynn waſſer eins ſpringen=
den brunnens / vñ

Abbildung 2.6: *Ausschnitt aus dem RIDGES-Korpus: Transkription von Zeilenumbrüchen, mit Erhaltung der Worttrennung und mit Markierung durch Trennungszeichen.*

Alle Transkriptionen übernehmen die Kennzeichnung einer durch einen Zeilenumbruch getrennten Wortform.[66] Die Markierung eines Zeilenumbruchs wird im Anselm-Korpus und im HIPKON-Korpus mit einem Doppelbindestrich, im RIDGES-Korpus mit schrägen Doppelbindestrich (⸗) übernommen. Die Trennung der Wörter selbst wird in HIPKON aufgehoben (*ge(=)heilet*). Anselm und RIDGES trennen die Wortteile in ihrer Transkription und unterscheiden sich aber in den Fällen, bei denen

[66] Wenn Textabschnitte nicht leserlich sind, wird das auch in Korpora markiert. Die Markierung im Anselm-Korpus wird mit `<...>` umgesetzt (Abbildung 2.4). Auch hier gibt es unterschiedliche Lösungen für dasselbe Problem. So markiert das RIDGES-Korpus dies mit _ .

Wörter auch aufgrund eines Zeilenumbruchs getrennt werden, aber keine Markierung vorhanden ist. Das Anselm-Korpus fügt ein *(=)* zusätzlich ein (*hier(=)* und *om*), das RIDGES-Korpus fügt keinerlei Markierung ein (*wur* und *tzel*). Die verschiedenen Ansätze greifen unterschiedlich stark interpretierend ein. Das Einfügen von Zeichen (hier Doppelbindestriche) ist beispielsweise eine Art Inline-Annotation (Ausweisung eines Zeilenumbruchs) und kann bereits als eine Art Normalisierung interpretiert werden. Die Unterscheidung zwischen Normalisierung und Transkription ist damit weniger kategorial und bewegt sich vielmehr auf einem Kontinuum, womit auch die jeweiligen Definitionen von (Primär-)Text variieren können.

Die Transkripte in jedem Korpus beziehen sich auch jeweils auf eine Komponente der Vier-Level-Beschreibung von Texten – also auf konzeptionell unterschiedliche Vorlagen. Eine mögliche Interpretation wäre für RIDGES und Anselm: Die Transkription in RIDGES bezieht sich auf einzelne Werke, wovon jeweils eine Expression und deren Manifestation von frühneuhochdeutschen Kräuterkundetexten transkribiert wird (historische gedruckte Bücher, Erstauflagen). Das Anselm-Korpus enthält hingegen Transkripte verschiedener Expressionen desselben Werks. Die Transkriptionen beziehen sich so möglicherweise auf unterschiedliche Textdefinitionen und sind selbst unterschiedlich stark unabhängig davon. Die Frage, inwieweit Transkriptionen eigene Expressionen besitzen oder Manifestationen darstellen, kann hiermit nicht beantwortet werden.

Gleiches gilt für die Normalisierungen; die Projekte erarbeiten jeweils unterschiedliche und aus dem Forschungsprozess heraus motivierte Kriterien, nach denen normalisiert werden soll; z. B. die jeweilige moderne orthographische Norm, wie sie im Deutschen der Duden (Dudenredaktion 2016) vorgibt. Die Normalisierung stellt damit eine weitere interpretative Text-Ebene im Korpus. Wie bereits angedeutet, kann die Normalisierung entweder eigenständig und getrennt von der Transkription durchgeführt werden wie bei RIDGES, oder sie ist ein Teil der Transkription und als Inline-Annotation vorhanden, wie sie auch Petrova et al. (2009) realisieren. Auch Parallelkorpora können pro Paralleltext eine Normalisierung erhalten (Dipper et al. 2015). Die Beispiele in Abbildung 2.7 und Abbildung 2.8 zeigen, wie unterschiedliche Perspektiven auf einen historischen Text in Form von Transkriptionen und Normalisierungen unter Einbezug der Korpusarchitektur realisiert werden können.

```
<lb n="8a,900,3003">
    <J IR="kop"><KON>und</KON></J>
    <J IR="kons" norm="dass" type="E" dir="V">
        <SUB>das</SUB></J>
    <subj real="Pron">wir</subj>
    <KOR>also</KOR> das Mal vor den Schweden <!--hier line-->
    <praed><V ID="Inf"><VV>bleiben</VV></V></praed>
    <praed><V ID="Fin"><MV>konten.</MV></V></praed></lb>
<line n="13"/>
```

Abbildung 2.7: *Ausschnitt aus dem Kasseler Junktionskorpus als Beispiel für eine Integration von verschiedenen Textkonzepten in einer SGML-Struktur. Bauernleben (1636-67).*

Das Kasseler Junktionskorpus (KAJUK)(Ágel und Hennig 2008; Hennig 2013b) ist in einer Standard Generalized Markup Language (SGML)-Struktur erstellt worden und enthält Elemente, die wiederum Attribute zugewiesen bekommen können. Der Ausschnitt aus dem KAJUK in Abbildung 2.7 besitzt nach den Annotationsrichtlinien[67] ein diplomatisches Transkript, realisiert als Zeichenkette, und eine normierte Schreibung, realisiert über das Attribut `@norm`, das in der dritten Zeile in Abbildung 2.7 dem Element `<J>` zugeordnet ist. Die historische Form des Belegs *das* ist mittels der Annotation über das Attribut `@norm` nach den Richtlinien der modernen Orthographie zu *dass* normalisiert und als Konjunktion annotiert. So wäre eine erste Interpretation folgende: Der historische Text liegt als Transkript in Plaintext vor, der wiederum annotiert worden ist. Die Normierung ist punktuell, aber getrennt vom historischen Text zugewiesen. Die Architektur von KAJUK erweist sich als komplex hinsichtlich der Identifikation der verschiedenen Arten von Text – Primärtext, Normalisierung, Annotation. Auch Ellipsen (Hennig 2013a) sind in historischer Schreibweise in das Transkript nachträglich rekonstruierend eingefügt und als solche mit dem Element `<E>` oder über ein Attribut wie `@type` wie in Abbildung 2.7, Zeile 3, annotiert. Die Konjunktion *das* ist also eine nachträglich zur historischen Vorlage hinzugefügte Worteinheit, die sich an der historischen Schreibung orientiert und in Form einer Annotation normalisiert wird. Damit verschwimmen die Grenzen zwischen historischem sprachlichen Material, Transkriptionen, Normalisierungen und weiteren Annotationen. De facto wurde eine artifizielle Wortform durch eine historisch angepasste Schreibung normalisiert (Abbildung 2.7).

[67] `http://www1.uni-giessen.de/kajuk/dokumentation.htm` (besucht am 03.08.2016).

dipl	von	Geiſ⸗	ſen	vnnd	Haſen	zuverſtehen	
clean	von	Geissen		vnnd	Hasen	zuverstehen	
norm	von	Geißen		und	Hasen	zu	verstehen

Abbildung 2.8: *Ausschnitt aus dem RIDGES-Korpus als Beispiel für eine Integration von verschiedenen Textkonzepten in multiplen Segmentierungen. PflantzGart (1639).*

Ein weiteres Beispiel für die Integration von Textkonzepten ist RIDGES (Abbildung 2.8)[68], das zwei verschiedene Normalisierungen in einer Korpusarchitektur vereint, die technisch gesehen als eigenständige und unabhängige Segmentierungen vom diplomatischen Transkript in die Korpusarchitektur eingebaut sind (Krause et al. 2012). Konzeptionell sind damit zwei Varianten für Normalisierungen – *clean* und *norm* – im Korpus integriert. Die historische Wortform *zuverſtehen* wird in der Ebene *clean* durch die Ersetzung von Sonderzeichen automatisch zu *zuverstehen* normalisiert und in der Ebene *norm* manuell auf Grundlage der modernen Orthographie durch *zu* und *verstehen* normalisiert. Weiterhin werden auch grafische Eigenschaften der historischen Vorlage in der *dipl* berücksichtigt. In der *dipl*-Ebene befinden sich zwei Einheiten *Geif⸗* und *fen*, die in der *clean*-Ebene als eine Einheit *Geissen* ohne ⸗ zusammengeführt werden. Hierbei handelt es sich um ein Wort in der historischen Vorlage, dass von einem Zeilenumbruch betroffen ist. Dies wird in *dipl* übernommen und in *clean* normalisiert.

Damit wird pro Textebene die Tokenisierung verändert. Die Ebenen *dipl* und *norm* besitzen in diesem Abschnitt zwar die gleiche Anzahl an Tokens, deren Werte sich aber unterscheiden. Die Ebene *clean* besitzt wiederum eine andere Anzahl an Tokens, mit wiederum zu *dipl* und *clean* unterschiedlichen Werten. So existieren technisch gesehen mehrere unabhängige Segmentierungen, *dipl*, *clean* und *norm*, auf denen jeweils annotiert werden kann. Normalisierungen können also ein Anwendungsfall für multiple Segmentierungen darstellen. Ein weiteres Beispiel dafür, welche Funktionen multiple Segmentierungen in historischen Korpora übernehmen können, zeigt das Anselm-Korpus, das die verschiedene Versionen desselben historischen Texts in einem Mehrebenenkorpus mit Hilfe multipler Segmentierung aligniert.

Aus diesen Überlegungen folgt, dass historische Korpora sehr stark in ihren verschiedenen Konzepten zu Text, Transkription/Normalisierung bzw. Annotation vari-

[68] RIDGES Version 5.0. PflantzGart_1639_Rhagor.Treffereferenzlink: `https://korpling.german.hu-berlin.de/annis3/?id=60e999b8-4953-4cf2-b3eb-789c019e5e07` (besucht am 01.02.2017).

ieren und verschiedene Korpusarchitekturen verwendet werden können. Es ist kaum möglich, einheitliche Textdefinitionen zwischen den hier betrachteten Korpora zu identifizieren. Mit dem jeweiligen Blickwinkel der fachspezifischen Forschungsfrage ändert sich unter Umständen die Interpretation von primären und sekundären Texten, der darauf basierenden Annotationen und deren Kategorien. Eine allgemeine Zuordnung der Transkription zum Vier-Level-Modell der FRBR ist so nicht ohne Weiteres möglich. Daher werden in dem hier erarbeiteten Modell Transkriptionen und Normalisierungen als Annotationen verstanden, die ihre jeweils spezifischen Annotationskategorien erhalten und in bestimmten Annotationskonzepten (und Formaten) umgesetzt werden. So muss die Korpusdokumentation nicht pro Korpus oder allgemein die Frage nach einer Definition von Text/Primärtext/Primärquelle beantworten und überlässt dies jeweils den Forschungsansätzen selbst (vgl. Kapitel 6).

2.7.3 Bearbeitung von Korpora am Beispiel von RIDGES

Der Forschungsdatenzyklus (Digital Curation Centre 2010) beschreibt die Erstellung, Bearbeitung, Weiterentwicklung und Veröffentlichung von Forschungsdaten ganz allgemein. Anhand des RIDGES-Korpus, das seit mehreren Jahren für korpuslinguistische Forschung verwendet wird, wird ein kurzes Beispiel für Bearbeitungsschritte innerhalb eines Forschungsdatenzyklus gegeben. Ziel ist es, daraus plausible Wiederverwendungsszenarien für textbasierte Korpora zu entwickeln, die in Kapitel 3 genau definiert werden.

Das RIDGES-Korpus ist in ein Projekt integriert, das seit mehreren Semestern mit Studierenden aus verschiedenen Bachelor- und Masterstudiengängen ein historisches Korpus des Frühneuhochdeutschen im Rahmen der forschungsorientierten Hochschullehre an der Humboldt-Universität zu Berlin erstellt.[69] Das RIDGES-Korpus wird so über die Jahre stetig weiterentwickelt und wiederverwendet.

Das Korpus ist 2011 in seiner ersten Version und 2016 in seiner bislang neuesten Version 5.0 frei verfügbar unter einer Creative Commons (CC)-Lizenz veröffentlicht.[70] Die Bearbeitungs- und Versionsgeschichte des Korpus ist umfangreich und soll für die hier vorliegende Arbeit daher als authentische Vorlage für Verwendungs- und Bearbeitungsszenarien eines Korpus fungieren. Die nachfolgende Tabelle 2.3

[69] Für eine Übersicht der Seminare vgl. `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/teaching-de` (besucht am 21.11.2016).

[70] `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download`. CC BY 3.0 DE, `https://creativecommons.org/licenses/by/3.0/de/`, (beide besucht am 04.08.2016).

listet grobe Parameter des Korpus auf, die sich über die verschiedenen Versionen geändert haben.

Tabelle 2.3: *RIDGES Herbology in den unterschiedlichen Versionen mit allgemeinen Angaben zur Größe und Tiefe des Korpus.*

Version	Token	Dok.	Eb.	Format
1.0	63.734	14	39	TEI, EXMARaLDA, PAULA, EXCEL, relANNIS
2.0	60720	13	42	TEI, EXMARaLDA, PAULA, EXCEL, relANNIS
3.0	122698	22	61	(TEI), EXMARaLDA, EXCEL, relANNIS
3.1	122698	22	61	EXMARaLDA, (EXCEL), relANNIS
4.0	153732	29	58	EXMARaLDA, EXCEL, relANNIS
4.1	154267	29	78	PAULA, EXCEL, relANNIS
5.0	182738	36	108	EXCEL, PAULA, ANNIS

Das Korpus wird in Seminaren um weitere Textauszüge erweitert und annotiert. Die Größe und die Tiefe des Korpus (Anzahl der Texte und Annotationen) variiert von Seminar zu Seminar. So gibt es Annotationsebenen, die in den ersten Versionen noch weiter annotiert, aber in späteren Versionen entweder ersetzt werden oder ersatzlos wegfallen. Einzelne Tags oder ganze Tagsets werden verändert, gelöscht oder erweitert. Auch bereits transkribierte Texte entfallen in späteren Versionen ersatzlos.

Tabelle 2.3 zeigt, dass über verschiedene feinkörnig definierte Versionsnummern die Anzahl der Dokumente, Annotationen und Formate zu- und abnehmen. Von Version 1.0 zu 2.0 fällt beispielsweise ein Dokument weg, womit das Sampling verändert ist.[71] In RIDGES ist ein Dokument ein transkribierter Text, der ein Auszug aus einem historischen Druck darstellt. Zum Teil aber existieren zwei oder mehr (unterschiedlich lange) Auszüge aus demselben Druck in unterschiedlichen Dokumenten. Die Dokumente variieren damit in der Größe und das Korpus besitzt ein spezielles Korpusdesign und eine eigene Dokumentdefinition, das sich auf die traditionelle Publikationsformen und -abschnitte eines Textes stützt.

Neben der sich verändernden Anzahl und Menge der Texte variieren die Annotationen ebenfalls: Es sind über alle Versionen hinweg Annotationsebenen dazugekommen, die das Korpus in seiner Tiefe anreichern. Wiederum wird aus einer solch einfachen Auflistung nicht ersichtlich, inwieweit konkrete Annotationsebenen von einer Version zu nächsten entfernt, ersetzt oder verändert sind.

Inhaltliche Anpassungen oder Erweiterungen der Annotationen können nur mit

[71] Vgl. die Liste der Textauszüge aus der Version 5.0: `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download/download-v5-de`, (besucht am 11.09.2016).

einer Aufschlüsselung der Richtlinien und der Anwendungsfälle aufgezeigt werden. So sind bei jeder neuen Version die Transkriptions- und Normalisierungsrichtlinien erweitert und um viele Beispielanwendungen ausführlicher gestaltet worden. Mögliche Änderungen und Korrekturen der Tagsets von einzelnen Annotationsebenen oder -werten müssen dann von einer Version eines Korpus zu einer anderen in einer Korpusdokumentation erfasst werden. In der Version 4.1 wurden beispielsweise die Annotationsebenen für eine Analyse von Kompositabildung (*komp*, *komp_orth*, *prot*, *attr_gen* und *strD*) hinzugefügt (vgl. Perlitz 2014). Die Annotationsebenen zu einzelnen Abschnitten (Ebenen *div*) sind lediglich in neun Dokumenten in der 1.0 Version annotiert worden. Damit sind auch nicht alle Annotationsebenen in allen Dokumenten ausgewiesen, so dass auch die Tiefe pro Dokument variiert. Auch die Metadaten (pro Dokument) wurden von Version zu Version erweitert.

Darüber hinaus werden auch über die verschiedenen Versionen hinweg unterschiedliche Formate genutzt. Es gibt Formate wie das TEI-XML, die in neueren Versionen nicht mehr verwendet werden. Bis zur bislang letzten Version 5.0 reduzieren sich die Formate, in denen das Korpus annotiert gespeichert und analysiert wird. Für die Auslesung der Formate werden dann auch unterschiedliche Tools für die Annotation (EXCEL, Partitur-Editor von EXMARaLDA), Analyse (Such- und Visualisierungstool ANNIS) und Konversion (ExcelAddIns, Konvertierungstool Pepper[72])[73] des Korpus benötigt.

Eine solche Aufstellung zeigt weder inhaltliche oder strukturelle Veränderungen des Korpus an, noch wird klar, warum die Anzahl der Formate über die Versionen hinweg abnehmen. Auf der korpuseigenen Dokumentationshomepage[74] werden die Neuerungen des Korpus ausführlich beschrieben. So wurde ab der Version 2.0 die Korpusarchitektur grundlegend verändert und die multiple Segmentierung in das Mehrebenenkorpus integriert.

Zusätzlich zur Entwicklung der Korpusarchitektur selbst verändert sich die Verarbeitungspipeline des Korpus.[75] Beispielsweise wird die Konvertierungspipeline in

[72] Dieses Konvertierungstool ist modulbasiert, so dass pro Ausgangsformat ein spezielles Modul zum Einlesen und pro Zielformat ein spezielles Modul für den Export genutzt werden muss.

[73] Vgl. `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/v4.1/conversion_to_annis.pdf` (besucht am 02.01.2017).

[74] Z. B. für die Version 5.0 `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/documentation/documentation-v5-de` (besucht am 04.08.2016).

[75] In welcher Funktion die jeweiligen Formate gebraucht und wie sie aufeinander abgebildet worden sind, geht aus einer einfachen Auflistung nicht hervor. Diese Art der Information ist ebenfalls in einer Fließtextform auf dieser Homepage dokumentiert. Vgl. für die Version

den Versionen bis einschließlich 4.0 über eine Konvertierung von TEI nach EXCEL, dann nach EXMARaLDA, dann nach PAULA und relANNIS in drei Schritten mit verschiedenen Tools durchgeführt. In den letzten Versionen werden die Dateien in EXCEL direkt nach PAULA und ANNIS in einem gemeinsamen Schritt mit dem Konvertierungstool Pepper (Zipser und Romary 2010) konvertiert.

Von Version 4.1 zu 5.0 fällt auch auf, dass ein Format umbenannt wurde, von *relANNIS* zu *ANNIS*. Die Umbenennung macht offensichtlich, was sonst typischerweise durch eine Versionsnummer angezeigt wird: das Format hat sich ebenfalls geändert.[76] Noch nicht weiter berücksichtigt sind die Tools selbst, die ebenfalls immer wieder überarbeitet und in neuen Versionen veröffentlicht werden. So kann die Version der genutzten Tools oder Formate von Korpusversion zu Korpusversion ebenfalls variieren.

Die Größe und Tiefe des Korpus, die Bestandteile der Korpusarchitektur, also die Tokenisierung, die Annotationskategorien und -konzepte und Metadaten, sowie die genutzten Formate und Tools können Veränderungen unterliegen: Alle können zum Teil unabhängig voneinander erweitert, verändert, gekürzt, entfernt, kontrolliert und neu gesampelt werden. Damit verändern sich die Eigenschaften, die stark auf den Forschungsprozess und die Forschungsfrage bezogen sind: korpuseigene Eigenschaften (Tokenisierung, Annotationskategorien, -konzepte, Metadaten) und korpusexterne Eigenschaften, die eher unabhängig vom Korpus definiert werden können (Format, Tool). Weiterhin kommen zu jeder Vergrößerung des Korpus Informationen zu den transkribierten Texten hinzu. Bislang ebenfalls nicht berücksichtigt sind die verantwortlichen Personen, die das Korpus erstellen, annotieren, verändern oder veröffentlichen. Informationen darüber zählen auch eher zu den korpusexternen Eigenschaften, die eine Korpusdokumentation enthalten kann.

Bei dieser Fülle an Informationen stellt sich grundsätzlich die Frage, wie allgemein oder wie spezifisch und wie ausführlich eine Korpusdokumentation für den Zweck der Wiederverwendung gestaltet werden muss. Für welche Wiederverwendungsszenarien muss eine Korpusdokumentation erstellt werden? Welche Nutzerprofile müssen beachtet werden? Welche Informationen sind für welche Szenarien und welche NutzerInnen relevant? Da alle historischen Korpora wie RIDGES diese Eigenschaften und

5.0 `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/documentation/documentation-v5-de` (besucht am 02.01.2017).

[76] Die ausführlichen Korpusdokumentation auf der Projekt-Homepage von RIDGES enthält viele Informationen, damit ein Korpus von Dritten zum Zweck der Wiederverwendung erschlossen werden kann. Diese Informationen besitzen ihre eigene korpuspezifische Struktur und Granularität. Wie eine einheitliche Korpusdokumentation gestaltet werden kann, wird an Beispielen bisheriger Ansätze in Kapitel 5 und an einem eigenen Ansatz in Kapitel 6 diskutiert.

einen Forschungsdatenzyklus besitzen, kann hieraus verallgemeinert werden, dass für den Korpustyp historisches Textkorpus gilt, dass die Größe und Tiefe genauso Veränderungen unterliegen kann wie die Tokenisierung, die Annotationskategorien und -konzepte, die Metadaten, die genutzten Formate und Tools sowie die verantwortlichen Personen. Kapitel 3 widmet sich diesen Fragen.

3 Wiederverwendung von Korpora

Die zentrale Motivation dieser Arbeit ist die Ermöglichung der Wiederverwendung von Forschungsdaten mit Hilfe von Metadaten. Die Metadaten von Forschungsdaten stellen die Grundlage für die Forschungsdatendokumentation und sind eine Voraussetzung für die Wiederverwendung von Forschungsdaten und damit Gegenstand der Modellierung und der technischen Umsetzungen in dieser Arbeit. Dieser Abschnitt beschreibt, was unter Wiederverwendung von historischen Korpora verstanden wird und wie daraus Wiederverwendungsszenarien abgeleitet werden. Maßgeblich werden diese aus den verschiedenen Versionen von historischen Korpora wie RIDGES abgeleitet. Wichtig ist, dass die Szenarien aus authentischen Beispielen abgeleitet werden, um sicher identifizieren zu können, was Modellierung und Metadaten leisten müssen.

3.1 Motivation

Der Wiederverwendung von Forschungsdaten geht ein Teilen – *Data Sharing* – derselben voraus. So müssen die Korpuserstellerinnen und -ersteller ihr Korpus mindestens einer dritten Partei in irgendeiner Weise übergeben, damit es geteilt ist. Das Teilen kann grundsätzlich persönlich, über eine digitale Plattform mit Zugangsbeschränkungen oder frei erfolgen.[77] Das Teilen von Daten verstehe ich nach Borgmann (2012: 1060) als eine Art der Datenveröffentlichung, die es Dritten ermöglicht, Daten nutzen zu können. Eine wesentliche Motivation für diesen Ansatz ist, dass dieselben sprachlichen Ressourcen innerhalb eines Fachs, aber auch überfachlich als Forschungsgrundlage dienen.

In dieser Arbeit wird davon ausgegangen, dass Forschungsdaten nur nachhaltig sein können, wenn sie auch wiederverwendet werden:

[77] Es gibt weitere ähnliche Ansätze, die auf weiteren und anderen Datentypen und deren spezifischen Kontexten einen Weg zum Teilen dieser Daten erarbeiten, z. B. in Form der internationalen Research Data Alliance (Berman et al. 2014), der Planung eines Forschungsdatenzentrums für eine größere Gruppe an Forschungsdatentypen (Buddenbohm et al. 2016) oder des integralen Datenmangementplans der Humboldt-Universität zu Berlin (Dreyer und Vollmer 2016). Weitere Ansätze, die sich mit der Beschreibung beziehungsweise Dokumentation von Forschungsdaten auseinandersetzen, werden in Kapitel 5 diskutiert.

> A resource will be used if it still exists, if it is usable, and if a user finds it relevant. (Simons und Bird 2008: 90)

Eine Dokumentation der Forschungsdaten muss genau dies berücksichtigen. Welche Eigenschaften eine Ressource besitzen muss, um eine so definierte Nachhaltigkeit von Forschungsdaten zu gewährleisten, zeigt die folgende Liste:

1. The resource must be extant.
2. The resource must be discoverable.
3. The resource must be available.
4. The resource must be interpretable.
5. The resource must be portable.
6. The resource must be relevant. (Simons und Bird 2008: 90)

Forschungsdaten müssen also für die hier entwickelten Wiederverwendungsszenarien vorhanden, auffindbar, zugänglich, interpretierbar, übertragbar und relevant sein. Wie bereits aufgezeigt wurde, ist die Frage der Relevanz von Forschungsdaten nur von den Forscherinnen und Forschern selbst beantwortbar, da Relevanz in diesem Kontext jeweils vor dem Hintergrund einer Forschungsfrage und einem Forschungsziel bewertet werden kann.

Metadaten können einen Teil dieser Eigenschaften gewährleisten (Kapitel 4). Nur so können überhaupt die Voraussetzungen für eine Wiederverwendung von Forschungsdaten geschaffen werden. Eine weitere Voraussetzung dazu ist, dass ein konkreteres Verständnis von Wiederverwendungsszenarien erarbeitet wird, was dieser Abschnitt der Arbeit leistet. Dafür notwendig ist ebenso eine überfachliche Perspektive auf die Forschungsdaten, wie sie in Kapitel 2 erarbeitet wurde.

Neben der konzeptionellen Möglichkeit, dass eine historische Ressource mehrere Forschungsfragen beantworten kann, gibt es eine projektplanerische Komponente, die überprüfen kann, vorhandene Korpora wiederzuverwenden. Es gibt bereits Beispiele für eine doppelte Digitalisierung derselben sprachlichen Ressource. Damit ist die Wiederverwendung von Forschungsdaten nicht nur eine theoretische Überlegung. Ein Beispiel für eine überfachliche doppelte Digitalisierung ist der Kräuterkundetext *Wundarznei*. So zeigen Leipold et al. (2015) anhand des Kräuterkundetexts *Wundarznei*, welche Aufgaben und Fragestellungen ein literaturwissenschaftliches Editionsvorhaben besitzt. Zentral sind hierbei die verschiedenen Handschriften des historischen Denkmals, die in der Edition berücksichtigt werden müssen. In dem

linguistischen Korpus RIDGES ist genau derselbe Kräuterkundetext (neben weiteren anderen) aufbereitet und digitalisiert, um die Entstehung und Entwicklung des wissenschaftlichen Registers im Deutschen zu untersuchen. Trotz der unterschiedlichen Forschungsfragen und Umsetzungen lassen sich auch über die Textwahl hinaus Parallelen feststellen: Die Projekte nutzen diplomatische Transkriptionen und Lemmatisierungen.

Aus diesem Beispiel wird bereits ersichtlich, dass eine Mehrfachverwendung von Forschungsdaten aus der Perspektive der Ressourcen plausibel ist. Ein Vorteil der Wiederverwendung derselben Ressource liegt projektplanerisch darin, dass Arbeits- sowie Forschungszeit in einem Projekt oder in der institutionellen Arbeit nicht auf die erneute Digitalisierung und vollständige Annotation verwendet werden muss. Die Erstellung einer Ressource ist kosten- und zeitaufwändig. Mit der Wiederverwendung von bereits vorhandenen Forschungsdaten kann ein Schwerpunkt eines Forschungsunternehmens entweder auf die Erstellung weiterer Ressourcen oder auf die Analyse und Auswertung der Daten gelegt werden. Zusätzlich müssen dann nicht zwangsläufig neue Datenmodelle und -formate erarbeitet werden, wenn schon vorhandene wiederverwendet werden können, was wiederum nur mit den entsprechenden fachlichen Kooperationen zu leisten wäre.

Das Teilen von Forschungsdaten ist in vielen Fachbereichen längst ein gängiges Verfahren, um bereits vorhandene Forschung zu reproduzieren oder zu verifizieren, Resultate zu publizieren und neue Forschung auf vorhandenen Daten zu ermöglichen (vgl. Borgmann 2012). Das Teilen der Daten ist eine Voraussetzung für deren Wiederverwendung. Daraus leiten sich einige der Wiederverwendungsszenarien in Abschnitt 3.2 ab.

Es ist prinzipiell möglich, eine doppelte Digitalisierung von Ressourcen zu verhindern, wenn es einerseits Kenntnis über andere Projekte gibt, die dieselben Texte digitalisieren, und wenn andererseits diese Digitalisate für Dritte dokumentiert sind und zur Verfügung stehen. (Meta-)Suchtools, Repositorien und Archive können die Suche nach Daten grundsätzlich ermöglichen (Abschnitt 2.6). Ohne eine einheitliche, zentrale und extensive Dokumentation mit Hilfe von Metadaten können Dritte in den jeweiligen Anwendungen die vorhandene Menge an Forschungsdaten nicht erschließen (Kapitel 4). Neben dritten Forscherinnen und Forschern sind es auch die initialen Korpuserstellerinnen und Korpusersteller, die ihre eigenen Forschungsdaten kurz-, mittel- oder langfristig wiederverwenden. So können die initialen Korpuserstellerinnen und -ersteller auf die ersten Forschungsfragen aufbauende, weitere und andere Fragestellungen an das von ihnen erstellte Korpus stellen. Abgeschlossene

oder noch laufende Korpusprojekte können in neue Projekte integriert oder mit gleichen und wechselnden Mitarbeiterinnen und Mitarbeitern weiterentwickelt werden. RIDGES ist z. B. in ein neues Projekt LangBank[78] eingebunden und parallel dazu weiterhin Gegenstand der universitären Lehre (vgl. Abschnitt 2.7.3).

Ob eine Wiederverwendung von Forschungsdaten auch inhaltlich sinnvoll oder eine aufgefundene Ressource relevant ist, entscheidet der Einzelfall beziehungsweise die Forschung selbst. Folgende Abschlussarbeiten der Humboldt-Universität zu Berlin im Fachbereich Linguistik sind Beispiele für die verschiedenen Wiederverwendungsmöglichkeiten von Korpora: Dietterle (2016) unternimmt eine umfassende Suche nach Belegen in deWaC über die Suchmaschine CQP der Humboldt-Universität zu Berlin für die Untersuchung von Binnenklammerung (z. B. *Buchstaben(-folgen)*). Dabei wird deWaC selbst nicht verändert, die vorhandenen Annotationen dienen als Analysegrundlage, die dann in weiteren Schritten aufbereitet und ausgewertet wird. Lehmann (2015) erweitert das Freiburger Korpus der Datenbank für Gesprochenes Deutsch (DGD)[79] mit Annotationen für eine Registerforschung (vgl. Biber und Conrad 2009) und Belz (2013) fügt in BeMaTaC neue Annotationen für eine kontrastive Disfluency-Analyse hinzu. Perlitz (2014) erweitert RIDGES ebenfalls um Annotationsebenen für ihre Untersuchung der diachronen Entwicklung von Komposita und nimmt zusätzlich noch Korrekturen an den bestehenden Annotationen vor. Unabhängig von Umfang und Motivation der Wiederverwendung ist für eine solche Entscheidung eine umfassende Dokumentation der Korpora unerlässlich. Wie eine solche Dokumentation aufgebaut ist und welche Inhalte sie transportieren muss, ist Gegenstand dieser Arbeit. Inwieweit dieser Ansatz langfristig Forscherinnen und Forschern eine Wiederverwendung von Korpora tatsächlich ermöglicht, hängt auch von weiteren Faktoren ab und kann nur mittelfristig überprüft werden (Kapitel 8).

3.2 Wiederverwendungsszenarien

Die Wiederverwendung von Korpora wird in dieser Arbeit als eine vielfältige Forschungsarbeit aufgefasst, die sich aus unterschiedlichen Szenarien und unterschiedlichen Akteurinnen und Akteure zusammensetzt. Aus dem in dieser Arbeit erarbeiteten Korpustyp – einem historischen textbasierten Korpus – dessen Eigenschaften und den genannten Beispielen für die (Wieder-)Verwendung und Änderungen in Form von immer neuen Versionen eines Korpus lassen sich folgende abstraktere Szenarien

[78] `http://sfs.uni-tuebingen.de/langbank/` (besucht am 04.01.2017).
[79] `http://agd.ids-mannheim.de/FR--_extern.shtml` (besucht am 04.01.2017)

ableiten:

Szenario 1 (Analyse)**.** Korpora können teilweise oder vollständig, manuell, semi-automatisch oder automatisch analysiert werden.
Szenario 2 (Korrektur)**.** Korpora bzw. ihre Annotationen können teilweise oder vollständig, manuell, semi-automatisch oder automatisch korrigiert werden.
Szenario 3 (Tiefenanreicherung)**.** Korpora bzw. ihre Dokumente können teilweise oder vollständig, manuell, semi-automatisch oder automatisch mit weiteren Annotationen angereichert werden.
Szenario 4 (Größenanreicherung)**.** Korpora können manuell, semi-automatisch oder automatisch mit weiterem sprachlichen Material angereichert werden.
Szenario 5 (Tiefenreduktion)**.** Korpora bzw. ihre Dokumente können teilweise oder vollständig, manuell, semi-automatisch oder automatisch durch die Reduktion von Annotationen verkleinert werden.
Szenario 6 (Größenreduktion)**.** Korpora können teilweise oder vollständig, manuell, semi-automatisch oder automatisch durch die Reduktion von sprachlichem Material verkleinert werden.
Szenario 7 (Konvertierung)**.** Korpora können teilweise oder vollständig, manuell, semi-automatisch oder automatisch in andere Formate, und damit in andere Datenmodelle mit unterschiedlich exakter Abbildung von Annotationskonzepten konvertiert werden.
Szenario 8 (Referenzierung)**.** Korpora können referenziert werden.

Eine Analyse (Szenario 1: Analyse) kann dabei ganz vielfältige Formen und Ziele besitzen: eine Replikation einer bereits durchgeführten Analyse, eine auf bestehende Analysen aufbauende Analyse oder eine komplett unabhängige Analyse. In Szenario 1 werden weiterhin alle Typen von Studien, wie z. B. quantitative und qualitative, korpusbasierte oder korpusgetriebene Studien eingeschlossen, da alle Analysen auf Annotationen basieren. Das Korpus wird bei diesem Szenario nicht selbst verändert.

Szenario 8 (Referenzierung) verändert das Korpus ebenfalls nicht. Mit Referenzierung kann eine Referenz in einer Publikation oder in einer Datenbank gemeint sein. Eine Referenzierung ändert sich typischerweise, sobald sich das Korpus ändert (Szenario 2, Szenario 3, Szenario 4, Szenario 5 und Szenario 6).

Die Anreicherung oder Reduktion von Annotationen und sprachlichem Material hingegen verändert das Korpus inhaltlich und gegebenenfalls auch architektonisch. Dies kann der Fall sein, wenn ein Korpus beispielsweise neue Annotationskonzepte erhält. Auch eine Konvertierung wie in Szenario 7 verändert das Korpus, und zwar

einerseits, indem neue Instanzen in einem anderen Format geschaffen werden, und andererseits, indem bei Konvertierungen typischerweise Informationen neu angeordnet werden oder verloren gehen können.

Diese Szenarien können einzeln und in Kombinationen auftreten. So werden bei Perlitz (2014) vorhandene Annotationen korrigiert und neue Annotationen hinzugefügt. Da das geänderte Korpus in ein Analyseformat konvertiert und dann erneut analysiert wird, besteht das komplette Wiederverwendungsszenario in diesem Fall aus den einzelnen Szenarien: Szenario 1, Szenario 2, Szenario 4, Szenario 7 und Szenario 8. Ähnlich verhält es sich bei Lehmann (2015). Dietterle (2016) nutzt nur Szenario 1 (Analyse). Möglich ist ebenfalls, dass ein Nebenprodukt von Szenario 1 (Analyse) weitere Kategorisierungen des Korpus Szenario 3 (Tiefenanreicherung) sind, die wiederum in dasselbe Korpus zurückgespielt oder als ein neues, eigenständiges Korpus erstellt und veröffentlicht werden können. Gemeint ist damit die getrennte, nicht wieder zusammenzuführende Entwicklung an einer Ressource, das Ergebnis sind dann zwei unabhängige Korpora.

Neben den eigentlichen Wiederverwendungsszenarien, also den Vorgängen mit und am Korpus, sind die Forscherinnen und Forscher zu unterscheiden, die diese ausführen. In dieser Arbeit werden die Forscherinnen und Forscher ganz allgemein mit *Akteurinnen* und *Akteure* bezeichnet, die sich in Bezug auf den Kenntnisstand unterscheiden:

> **Akteurin/Akteur 1** (Initialerstellung)**.** Forschende erstellen ein Korpus.
> **Akteurin/Akteur 2** (Initialbearbeitung)**.** Forschende bearbeiten oder verwenden das von ihnen erstellte Korpus wieder.
> **Akteurin/Akteur 3** (Drittbearbeitung)**.** Forschende bearbeiten oder verwenden ein Korpus, das nicht von ihnen selbst erstellt oder bearbeitet worden ist.
> **Akteurin/Akteur 4** (Inital- und Drittbearbeitung)**.** Forschende bearbeiten oder verwenden das von Ihnen erstellte Korpus mit Dritten.

Akteurinnen und Akteure können einzelne Personen oder auch Personengruppen sowie Projekte und Institutionen sein. Diese Akteurinnen und Akteure sind vergleichbar mit jenen Gruppen, wie sie bei Simons und Bird (2008: 91) beschrieben werden, nämlich mit *Creators*, die Forschungsdaten erstellen, oder *User*, also Personen, die das Korpus benutzen wollen. Daneben gibt es andere Akteurinnen und Akteure wie beispielsweise Institutionen, die Forschungsdaten kuratieren. Letztere werden in dieser Arbeit nicht weiter berücksichtigt. Die hier vorgestellte Klassifika-

tion unterscheidet die Personen, die an Korpora arbeiten, noch genauer. So können *User* Akteurin/Akteur 1 (Initialerstellung), Akteurin/Akteur 2 (Initialbearbeitung), Akteurin/Akteur 3 (Drittbearbeitung) und Akteurin/Akteur 4 (Initial- und Drittbearbeitung) sein. Akteurin/Akteur 1 hingegen sind zugleich auch *Creator*.

Wenn verschiedene Akteurinnen und Akteure unterschiedliche Szenarien ausführen können, dann ist es auch möglich, dass zwei getrennte Entwicklungen auf demselben Korpus stattfinden. Ein solches Beispiel ist RIDGES: In Version 4.1 wurde das Korpus in seiner Größe erweitert – Szenario 4 – durch eine Seminargruppe – Akteurin/Akteur 3 – und mit weiteren Annotationen – Szenario 3 – durch weitere initiale Akteurinnen und Akteure – Akteurin/Akteur 2 – angereichert. Beide Entwicklungen sind anschließend in einer gemeinsamen Korpusarchitektur zusammengeführt worden. Dies muss nicht immer der Fall sein. Unabhängige Entwicklungen an einem Korpus durch verschiedene Akteurinnen und Akteure sind ebenfalls möglich (vgl. z. B. Dumont 2016).

In dieser Betrachtung von Wiederverwendungsszenarien wird jede Version eines Korpus als Menge aus verschiedenen Szenarien mit den dazugehörigen Akteurinnen und Akteuren verstanden. Die Szenarien werden als abgeschlossene Handlungen verstanden, die ein Resultat erzeugen, das als solches mit Metadaten beschrieben werden soll. Inwiefern Szenario 1 oder Szenario 3 in sich Prozesse mit einzelnen Schritten und Zyklen darstellen, wird hier nicht weiter betrachtet.

3.3 Ansatz zur Unterstützung der Wiederverwendung von Forschungsdaten

In dieser Arbeit wird ein Vorschlag zur Dokumentation von historischen Korpora zum Zweck der Wiederverwendung erarbeitet. Ein ‚information retrieval system' (vgl. Kapitel 4) soll mit Hilfe des hier vorgeschlagenen Metamodells für Korpusmetadaten und dessen Instanziierung Forscherinnen und Forschern ermöglichen, Korpusdaten aus einer Menge an Korpora zum Zweck der Wiederverwendung zu finden und alle relevanten Informationen für diesen Zweck über ein Korpus zu erhalten. Dieser Vorschlag fokussiert sich auf die Entwicklung eines Metamodells für Korpusmetadaten, welches die relevanten Korpuseigenschaften für die oben genannten Szenarien, Akteurinnen und Akteure identifiziert. Dieses Metamodell ist in ein Modellsystem eingebettet, um eine technische Umsetzung zu ermöglichen.

Weitere notwendige Ansätze für die Unterstützung einer inner- oder überfachli-

chen Wiederverwendung von Forschungsdaten sind Datenmodelle, Formate und Bearbeitungstools. So ermöglichen Annotationswerkzeuge wie EXMARaLDA (Schmidt und Wörner 2009), ELAN (Wittenburg et al. 2006) und ATOMIC (Druskat et al. 2014) Szenario 2, Szenario 3, Szenario 4, Szenario 5 und Szenario 6. Datenmodelle, Konvertierungspipelines und Frameworks wie PEPPER (Zipser und Romary 2010) unterstützen Szenario 7. Such- und Visualisierungstools unterstützen Szenario 1.

Diese Arbeit leistet einen Beitrag zur Wiederverwendung von Korpora mittels eines extensiven, einheitlichen Metamodells inklusive seiner Implementierung. Eine der größten Herausforderungen für diesen Ansatz ist folgende:

> As information seekers, people are generally most interested in resources as works and expressions, that is, in their information content. Nevertheless, they may, for a variety of reasons, prefer a particular manifestation of a work, or even a particular item. (Hider 2012: 20)

Das Anwenderszenario für diese Arbeit ist also folgendes: Es existiert eine Menge an unterschiedlichen historischen textbasierten Korpora aus verschiedenen Fächern, die an einem Speicherort liegen oder die über einen Metadatenzugriff zugänglich sind. Diese Menge wird einheitlich mit Metadaten beschrieben. Die Forscherinnen und Forscher kennen unter Umständen die Menge an Korpora nicht, kennen aber ggf. einzelne Korpora Dritter oder ihre eigenen aus dieser Menge. Die unterschiedlichen Forschergruppen – Akteurinnen und Akteure – suchen zum Zweck der Wiederverwendung – Szenarien – nach einem oder mehreren Korpora mittels Metadaten. Wenn Korpora gefunden werden, können diese mittels ihrer Metadaten weiter erschlossen und auf ihre Eignung überprüft werden.

Die Szenarien beschreiben, welche Ziele Akteurinnen und Akteure verfolgen. Sie sind entweder auf der Suche nach Korpora, die für ihre angestrebten Szenarien in Betracht kommen oder sie haben bereits Korpora vorliegen. Der in dieser Arbeit verfolgte Ansatz ist, dass Metadaten helfen können, relevante Eigenschaften der Korpora zu erschließen, ohne dass eine direkte Interaktion mit den initialen ErstellerInnen oder ein Einlesen und Manipulieren der Daten in einem ersten Schritt notwendig ist. In Kombination mit den in Kapitel 2 beschriebenen Eigenschaften von Korpora werden dann die relevanten Metadaten zu diesen Eigenschaften modelliert. Welche konkrete Strukturen können helfen, Korpora über Metadaten für Dritte erschließbar zu machen? Wie können diese in Form von Metadaten abgebildet werden? Wie Metadaten in dieser Arbeit definiert werden, welche Funktionen und Aufgaben

sie übernehmen können und wie sie als interdisziplinärer Zugang in dieser Arbeit verstanden werden, zeigt Kapitel 4.

4 Metadaten

Nachdem die Eigenschaften von historischen Korpora herausgestellt und deren Wiederverwendungsszenarien abgeleitet und generalisiert worden sind, wird nun der Begriff der `Metadaten` in Bezug auf diese Forschungsdaten und für den Zweck der Wiederverwendung definiert und klassifiziert.

4.1 Einordnung des Begriffs

`Metadaten` werden oft ganz allgemein mit 'Daten über Daten' beschrieben. Metadaten sind

> [...] structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information. (NISO 2004: 1)

Metadaten sind also Informationen, im weiteren Sinne Wissen oder Wissensrepräsentationen, die in irgendeiner Form strukturiert sind. Durch Metadaten wird es dann ermöglicht, die eben beschriebenen Daten oder Informationen zu organisieren, zu gebrauchen oder abzurufen.

> It [metadata] can be used to describe highly structured resources or unstructured information such as text documents. Metadata can be applied to description of electronic resources; digital data (including digital images); and to printed documents such as books, journals and reports. Metadata can be embedded within the information resource (as is often the case with web resources) or it can be held separately in a database. (Haynes 2004: 8)

Metadaten können demnach dafür benutzt werden, strukturierte und nicht strukturierte Daten zu beschreiben. Solche Daten sind z. B. elektronische Ressourcen oder

gedruckte Dokumente wie Bücher oder Zeitschriften. Die Metadaten liegen dann entweder eingebettet in der Ressource oder separat zur Ressource vor. So gibt es aus definitorischer Sicht keine Einschränkungen, worauf sich Metadaten beziehen können: die in Kapitel 2 genannten historischen Korpora können wie alle anderen Arten von Forschungsdaten mit Metadaten beschrieben werden. `Metadaten` werden in dieser Arbeit nach Hider (2012: 4) als „information resource descriptions“ verstanden: Neben Forschungsdaten können Metadaten auch Informationen beschreiben. In dieser Arbeit wird daher zwischen zwei Arten von Informationen bzw. Eigenschaften von Korpora unterschieden: Informationen können Eigenschaften des Korpus sein, die das Korpus selbst betreffen (`korpuseigen`), oder Eigenschaften darstellen, die andere, dritte Entitäten beschreiben, die unabhängig vom Korpus existieren (`korpusextern`). Neben den Eigenschaften der eigentlichen Korpusdaten können Korpusdokumentationen so auch korpusexterne Informationen enthalten. In dieser Arbeit werden Metadaten als Informationen über digitale Korpusdaten und Informationen zu diesen verstanden, die separat von der Ressource liegen (können) und Handlungen mit diesen ermöglichen.

Folgende Fragen sind dabei zentral und werden in den nachfolgenden Abschnitten beantwortet:

1. Wie können Metadaten die Eigenschaften der Korpora beschreiben?
2. Welche Funktionen können die Metadaten übernehmen?
3. Welche Handlungen sollen durch die Metadaten ermöglicht werden?
4. In welcher Form müssen die Metadaten erstellt werden?

Metadaten definieren sich in Abhängigkeit von dem, was sie beschreiben – also je nach Ressource oder Information über die Ressource. Dieser Objektbezug wird in Abschnitt 4.2 näher ausgeführt. Metadaten sind multifunktional und können nach unterschiedlichen Funktionen klassifiziert werden (Abschnitt 4.3). Ein wesentlicher Aspekt für Metadaten ist ihr zeitlicher Bezug, der eng mit dem Forschungsdatenzyklus diskutiert und in dieser Arbeit entwickelt wird (Abschnitt 4.4). Wichtig ist dabei z. B., welcher Zeitpunkt oder Zeitraum des Forschungsdatenzyklus einer Ressource beschrieben werden soll. Metadaten können verschiedene Handlungsszenarien mit einer Ressource ermöglichen (Abschnitt 4.5). Diese werden zusammen mit den in dieser Arbeit erarbeiteten Wiederverwendungsszenarien (Kapitel 3) diskutiert,

um herauszufinden, welche Handlungen auf der Basis von Metadaten Voraussetzungen für welche Wiederverwendungsszenarien darstellen können. Metadaten besitzen auch unterschiedliche Formen, die in Abschnitt 4.6 kurz dargelegt werden. Schließlich können Metadaten unter Erfolgskriterien dieser verschiedenen Handlungsszenarien evaluiert sowie unter verschiedenen Qualitätsmerkmalen bewertet werden (Abschnitt 4.7).

4.2 Objektbezug

Metadaten stehen im direkten Bezug zu dem, was sie beschreiben. Es ist wesentlich zu verstehen, welches Objekt, welcher Teil oder welche Ebene des Objektes mit Hilfe von Metadaten beschrieben wird (Haynes 2004: 67). Jedes Objekt oder jeder Teil eines Objektes besitzt andere Eigenschaften, die durch die Metadaten beschreibbar gemacht werden. In Kapitel 2 wurde beispielsweise gezeigt, dass ein Korpus durch seine Tokenisierung, Annotationsebenen und -kategorien beschrieben werden kann. Weiterhin ändern sich Metadaten, wenn sich das zu beschreibende Objekt bzw. dessen beschriebene Eigenschaften ändern. Die Eigenschaften einer Annotationsebene können beispielsweise durch Metadaten beschrieben werden. Die Metadaten einer Annotationsebene können die enthalten Werte (Annotationskategorien) sein. Wenn die Werte der Annotationsebene sich über die Zeit ändern, ändern sich auch die Metadaten, die die Annotation beschreiben.

Dieser Bezug zu den Daten wird in dieser Arbeit `Objektbezug` genannt. Das Objekt, was beschrieben werden soll, ist in Abschnitt 2.1 bereits ausführlich dargestellt worden: historische textbasierte Korpora, die sich durch spezifische korpuseigene und korpusexterne Eigenschaften auszeichnen. Die ausführliche Darstellung der verschiedenen Korpora in Kapitel 2 ist notwendig, um die Eigenschaften historischer Korpora zu identifizieren.

Metadaten können separat von dem Objekt, dass sie beschreiben, erstellt werden, da sie nicht selbstverständlich konzeptioneller oder technischer Teil dieses Objektes sein müssen. Metadaten können je nach Perspektive auch als Forschungsdaten reanalysiert werden. Metadaten können damit post hoc erzeugt oder definiert werden, weil sie ein Objekt erst nach (oder während) seiner Erstellung beschreiben können, beziehungsweise Daten nach ihrer Erstellung als Metadaten uminterpretiert werden können. In Bezug auf die Beziehung zwischen historischen textbasierten Korpora und deren Metadaten wird hier davon ausgegangen, dass Metadaten *konstruiert* sind (vgl. Coyle 2005; Miller 2011), da sie nicht Teil der natürlichsprachigen histo-

rischen Äußerung oder der Annotationen im Korpus sind. Metadaten von Korpora werden entweder in einem Erstellungsschritt[80] oder in einem nachgelagerten Schritt wie der Veröffentlichung des Korpus erstellt; beispielsweise in Form eines technischen Berichtes[81], in einem eigenen Metadatenformat[82] oder einer Homepage[83].

Wie Kapitel 2 gezeigt hat, sind Annotationen immer auch Interpretationen und sind an theoretischen Modellen, Sprachen und Begriffen gebunden. Die Entscheidungen, welche Annotationen wie für welche Analyse genutzt werden, werden für dritte Forscherinnen und Forscher idealerweise in Form von Fließtext in einem Zeitschriftenartikel oder Annotationsguidelines nachvollziehbar erklärt.

Um die Erstellung von Metadaten zu einem Objekt ebenfalls nachvollziehbar zu gestalten, bedarf es ebenfalls einer Art Fachtext. Metadaten werden aus einer forschungsorientierten Perspektive erstellt. Metadaten sind dann ebenfalls Interpretationen und lassen damit eine Metadokumentation, die die Metadaten und deren Funktionen einordnet und motiviert, unverzichtbar erscheinen. Es wird in Kapitel 5 herausgearbeitet, dass vorhandene Metadatenschemata eine solche Metadokumentation für historische Korpora für den Zweck der Wiederverwendung nicht oder nicht vollständig besitzen. Daher wird ein Metamodell für Korpusmetadaten im Rahmen dieser Arbeit entwickelt, das beschreibt, welche Metadaten für historische textbasierte Korpusdaten unter einer bestimmten funktionalen und formalen Perspektive notwendig sind. Damit sind Metadaten immer noch Interpretationen, sie werden aber aus einer empirischen Grundlage abstrakt nachvollziehbar hergeleitet, motiviert und können in verschiedenen Metadatenschemata realisiert werden (Kapitel 6 und Kapitel 7).

4.3 Funktionale Klassifikation

Metadaten sind *konstruktiv*, *kommunikativ*, *multifunktional* und *ausführbar*. Sie können helfen, Ressourcen zu identifizieren, zu archivieren, zu suchen oder zu strukturieren (Haynes 2004: 17), und sind damit *konstruktiv*. Forschungsdaten werden

[80] Einige Annotationswerkzeuge bieten eine Metadatenerfassung mit an wie z. B. EXMARaLDA (Schmidt und Wörner 2009), Bearbeitungspipelines wie Unstructured Information Management Application (UIMA), (Ferrucci und Lally 2004) oder Weblicht (Kok et al. 2015) geben Metadaten automatisch mit aus.

[81] Wie z. B. Annotationsrichtlinien für das historische Korpus RIDGES (Belz et al. 2016) oder für Falko (Reznicek et al. 2012).

[82] Wie es die TEI (TEI Consortium 2015) in ihrem TEI-XML-Format direkt mit anbietet oder wie separate Metadatenformate wie Component MetaData Infrastructure (CMDI).

[83] Wie z. B. KAJUK `http://www.uni-giessen.de/kajuk/index.htm`, GerManC `http://www.llc.manchester.ac.uk/research/projects/germanc/` (besucht am 04.08.2016).

für eine bestimmte Gruppe von Adressaten erstellt (Owens 2011). Vergleichbar mit Forschungsdaten werden auch Metadaten von Forschungsdaten für eine bestimmte Gruppe von Adressaten erstellt (Coyle 2005: 160).[84] Metadaten können an die Korpuserstellerinnen und -ersteller (Akteurin/Akteur 2), an andere, dritte Forschende (Akteurin/Akteur 3) oder an andere Metadaten adressiert werden. Das Wissen und die Erwartungshaltung der Metadatenerstellerinnen und -ersteller und deren Nutzerinnen und Nutzer nehmen dann jeweils Einfluss auf die Metadaten. So entsteht eine Nutzer- und eine Erstellerperspektive. Damit sind Metadaten nicht nur objektbezogen, sondern auch adressatenbezogen und somit *kommunikativ* (Miller 2011; NISO 2004). Metadaten sind *multifunktional*, weil sie vielfältig eingesetzt werden und unterschiedliche Zwecke erfüllen können (Haynes 2004; Miller 2011; NISO 2004). Metadaten sind in Abhängigkeit von ihrer Form und ihren Zweck dann *ausführbar*, da auf ihrer Grundlage `Handlungen` (vgl. Abschnitt 4.5) automatisch oder manuell ausgeführt werden können (Coyle 2005).

Häufig wird zwischen drei großen Klassen von Metadaten, die die oben genannten Funktionen erfüllen, unterschieden: `Deskriptive` Metadaten, `administrative` Metadaten und `strukturelle` Metadaten (Miller 2011: 12; NISO 2004: 1).

> 1. *Descriptive metadata* describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
> 2. *Structural metadata* indicates how compound objects are put together, for example, how pages are ordered to form chapters.
> 3. *Administrative metadata* provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it. There are several subsets of administrative data; two that sometimes are listed as separate metadata types are:
>
>   a) *Rights management metadata*, which deals with intellectual property rights, and

[84] Neben den Anwendungen zur Datenspeicherung und -archivierung werden Metadaten auch zur Analyse und -visualisierung eingesetzt, wie die Sketch Engine für Lexikographie (Kilgarriff et al. 2014), die Penn Treebank (PTB) für Baumbanken (Kroch und Taylor 2000), oder für Mehrebenenkorpora in ANNIS (Krause und Zeldes 2016). Die Metadaten und ggf. deren Schematisierung können entweder direkt in dem Annotationsformat abgelegt, separat gespeichert oder später beim Import in ein Analysetool hinzugefügt werden. Häufig werden die Metadaten in freien Attributwertpaaren angegeben, die eigentliche Modellierung der Metadaten erfolgt über die Datenformate der Forschungsdaten.

b) *Preservation metadata*, which contains information needed to archive and preserve a resource.

(NISO 2004: 1)

Deskriptive Metadaten beschreiben den Inhalt der Ressource näher und geben beispielsweise den Titel, den Autor, die Textsorte, eine Zusammenfassung oder das Erscheinungsjahr an; sie liefern eine Art intellektuellen Zugang zur Ressource (Miller 2011: 10). Administrative Metadaten helfen die Ressource zu managen, in dem sie technische Informationen über beispielsweise die Erstellung und den Datentyp angeben. Hier besteht dann ein enger Bezug zum Lebenszyklus von Forschungsdaten. Administrative Metadaten zum Rechtemanagement geben unter anderem Lizenzen oder verantwortliche Institutionen an. Weiterhin gibt es administrative Metadaten, die der Erhaltung der Ressource dienen können. Strukturelle Metadaten erfassen den Aufbau und Zusammenhang der Ressource wie die Auszughaftigkeit der Ressource[85] oder verschiedene visuelle Ansichten einer Ressource.

Je nach Perspektive gibt es noch feinere Unterscheidung, die eine bestimmte Funktion stärker herausstellen. So können auch `technische Metadaten` und Metadaten zum Gebrauch der Ressource (`Use`) separat definiert werden (vgl. z. B. Gilliland 2008; Haynes 2004), die bei NISO (2004) unter die administrativen Metadaten eingeordnet werden. Unter technischen Metadaten werden Informationen zusammengefasst, die die technische Umsetzung – beispielsweise Format und Software – betreffen. Metadaten zum Gebrauch der Ressource beinhalten Informationen über die Nutzung von Visualisierungen oder Analysemethoden für die Ressource. Diese Metadaten können neben dem Objektbezug auch einen zeitlichen Bezug besitzen (vgl. Abschnitt 4.4).

Die jeweilige Gewichtung und der benötigte Umfang der Metadaten hängt, wie oben beschrieben, vom Zweck der Dokumentation, der jeweiligen Ressource (Objekt) sowie den Adressaten ab. Diese Typen von Metadaten sind durch ihre inhaltlich-beschreibende Perspektive gekennzeichnet.

In Bezug auf Korpusmetadaten können deskriptive Metadaten beschreiben, welcher Text digitalisiert und annotiert ist. Unter den strukturellen Metadaten wird beispielsweise beschrieben, ob es sich um einen Textauszug oder um eine komplette Publikation handelt. Die administrativen Metadaten können die Korpuslizenz und die technischen Metadaten die verwendeten Annotationsformate beschreiben.

Diese Klassifikation nimmt eine eher inhaltliche Perspektive ein, indem sie die

[85] Darunter wird verstanden, ob die gesamte Ressource, ein Auszug oder mehrere Auszüge oder Varianten davon gemeint sind.

Metadaten nach der Art der Information über die Ressource einordnet. Aber erst eine weitere zweckbezogen Einordnung von Metadaten kann helfen zu systematisieren, welche Eigenschaften in Form von gerade diesen inhaltlichen Metadaten benötigt werden, um eine Ressource für ein bestimmten Zweck zu beschreiben. In dieser Arbeit werden daher Metadaten über zwei Perspektiven klassifiziert: Eine inhaltlich-funktionale, die klassifiziert, welcher Art von Information vorliegt und wie sie strukturiert ist; und eine zweite zweckgezogene Perspektive, die klassifiziert, welche Informationen für einen bestimmten Zweck relevant genutzt werden sollen. So können dieselben Metadaten unterschiedliche Zwecke besitzen und unterschiedliche Handlungen unterstützen (Haynes 2004: 17 und Abschnitt 4.5).

4.4 Zeitlicher Bezug

Für die hier vorliegende Arbeit ist eine weitere, dritte Perspektive neben der inhaltlich-funktionalen und zweckbezogenen notwendig: die zeitliche Perspektive. Gerade da sich die Erschließung eines Korpus zum Zweck der Wiederverwendung stark auf die deskriptiven und technischen Metadaten eines Korpus stützt, ist nicht nur relevant, welche Metadaten jeweils konkret angegeben werden, sondern auch, welche zeitliche Perspektive diese einnehmen. Es ist also relevant, ob die Metadaten Korpuseigenschaften eines gesamten Zeitraums oder eines Zeitpunkts abbilden.

Das folgende abstrakte Beispiel soll zeigen, was unter der zeitlichen Perspektive von Metadaten in dieser Arbeit verstanden wird. Dieser zeitliche Bezug steht in enger Verbindung mit dem Forschungsdatenzyklus von Korpora. Nehmen wir folgendes Beispiel einer Korpusbearbeitung an: Ein Korpus wird annotiert. Dabei werden die Bearbeitungsschritte 1-7 innerhalb eines bestimmten Zeitraums getätigt (Abbildung 4.1). Diese Bearbeitungsschritte sind Annotationen der Tags A, B und C. In dieser Annotationsebene werden nacheinander der Tag A annotiert, ein nächster Tag B hinzugefügt, der Tag A gelöscht und der Tag B durch ein Tag C ersetzt.

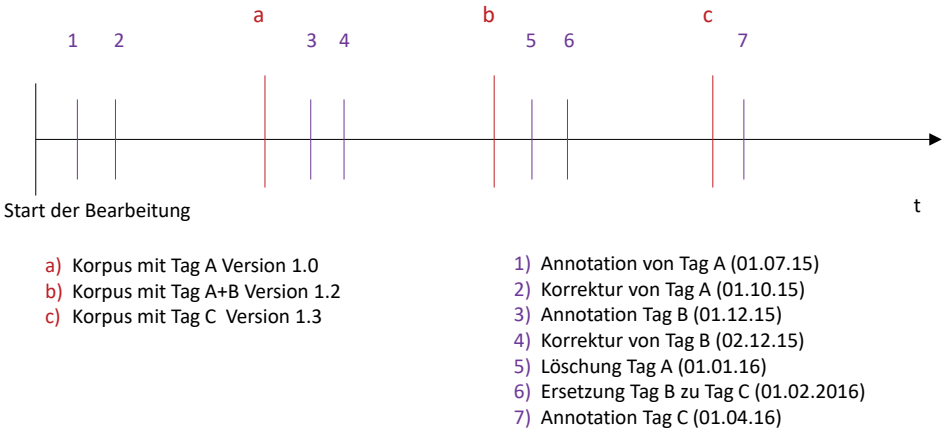

Abbildung 4.1: *Bearbeitungsschritte einer Annotationsebene in einem Korpus (1-7) und die Zeitpunkte der Korpusdokumentation (a-c). Ein Beispiel für die zeitliche Perspektiven von Metadaten.*

Zwischen diesen Bearbeitungsschritten 1 bis 7 wird dieses Korpus dreimal veröffentlicht, nämlich zu den Zeitpunkten a) bis c). Diese drei Veröffentlichungen sind dann jeweils Versionen eines Korpus. Eine Version eines Korpus ist ein Produkt des Forschungsdatenzyklus. Zu den verschiedenen Zeitpunkten a) bis c) hat sich jeweils die Annotationsrichtlinie geändert, die in der Korpusdokumentation mit Metadaten beschrieben werden soll. Die erste Version das Korpus zum Zeitpunkt a) enthält nur den Tag A, die zweite die Tags A und B und die dritte nur den Tag C. Welche Eigenschaften der Annotationsebene eines Korpus sollen zu einem oder mehreren Zeitpunkten oder Zeiträumen mit Metadaten beschrieben werden?

Metadaten können exhaustiv die Bearbeitungsschritte in der Liste 1 bis 7 beschreiben und damit den Erstellungsvorgang – also den Verlauf oder Prozess von Annotationsschritten – im Detail mit Zeitstempeln auch unabhängig von den Versionen eines Korpus dokumentieren. Diese Art von Metadaten können als Prozessmetadaten bezeichnet werden. Ein Beispiel zu solchen prozessorientierten Metadaten gibt K. Eckart (2015). Prozessmetadaten können darüber hinaus auch die zeitliche Entwicklung oder Performanz von Tools wie einem Parser (in verschiedenen Versionen)

auf verschiedene Ressourcen abbilden. So kann z. B. festgestellt werden, wie gut ein Parser auf einer Ressource angewendet werden kann. Diese Daten, der Output des Tools, müssen hingegen nicht zwangsläufig als ein gespeichertes, zugängliches, nicht flüchtiges Produkt oder Ergebnis (Korpus) vorliegen. Damit kann der Schritt des Parsens der Daten in diesem Fall nicht Teil eines Forschungsdatenzyklus eines Korpus verstanden werden. Bei einem solchem Szenario (Testen eines Parsers auf Daten) ist nicht zwingend „ein letzter Schritt“ definiert. Sie sind eine Bearbeitungsdokumentation, die häufig eher im Zusammenhang mit dem Annotationstool nicht so sehr mit dem Korpus selbst stehen kann.

Typischerweise muss ein Korpus gespeichert und veröffentlicht werden, um es für Dritte zur Wiederverwendung zur Verfügung zu stellen. Korpora erhalten dann meist je Veröffentlichung eine Version mit Nummer. Das Korpus wird dann mit den Annotationen, die es konkret zu einem Zeitpunkt besitzt, veröffentlicht. Korpora besitzen ihren eigenen Forschungsdatenzyklus und werden aus mehren Bearbeitungsschritten erzeugt. Diese Bearbeitungsschritte können in Relation zum Veröffentlichungszeitpunkt permanent oder flüchtig sein. In Abbildung 4.1 ist der Bearbeitungsschritt 1 – die Annotation von Tag A – zu einem späteren Zeitpunkt in der Version c) in den Korpusdaten nicht mehr nachvollziehbar. Der Bearbeitungsschritt 6 – die Ersetzung von Tag B zu Tag C – ist zu demselben Zeitpunkt in der Version c) noch erkennbar. Die Beschreibung desselben Korpus kann also auch von einem Zeitpunkt ausgehen und enthält dann in Abhängigkeit vom jeweiligen Zeitpunkt andere Informationen.

Nehmen wir die drei Zeitpunkte a), b) und c) (Abbildung 4.1), zu denen das Korpus jeweils unabhängig voneinander mit Korpusmetadaten beschrieben werden kann. Es gibt einen Zeitpunkt a), an dem das Korpus mit der Annotation von Tag A vorliegt. Bis zu diesem Zeitpunkt wurde Tag A annotiert und korrigiert. Zu einem Zeitpunkt b) kann beschrieben werden, dass das Korpus die Tags A und B erhält. In Bezug auf die Wiederverwendungsszenarien ist dann relevant zu dokumentieren, dass bis zu diesem Zeitpunkt eine Version eines Korpus bestimmte Tags enthält, deren Ausweisung kontrolliert wurde.

Zu einem Zeitpunkt c), an dem das Korpus vorliegt, wird jedoch gerade nicht dokumentiert, dass es zu einem früheren Zeitpunkt die Tags A und B im Korpus gab, da diese zu Zeitpunkt c) nicht mehr im Korpus nachvollziehbar sind und damit für eine weitere Analyse oder Bearbeitung des Korpus nicht mehr berücksichtigt werden können. Für Zeitpunkt c) wird ebenfalls nicht dokumentiert, dass das Korpus weitere Annotationen erhält beziehungsweise in Zukunft erhalten wird.

Wenn das Korpus zu den Zeitpunkten a)-c) jeweils als eine neue Version gespei-

chert, nicht flüchtig und zugänglich vorliegt, dann können die Dokumentationen der jeweiligen Versionen Aufschluss darüber geben, welche Bearbeitungsschritte für eine der jeweiligen vorausgegangenen Versionen notwendig war. Die Metadaten beziehen sich dann jeweils auf ein festgelegtes Produkt oder Ergebnis aus mehreren Bearbeitungsschritten (Forschungsdatenzyklus), das als konkrete Version(en) eines Korpus verstanden werden kann.

Die Konzentration auf eine konkrete zeitliche Perspektive ermöglicht es, die Auswahl von Informationen, die Metadaten über ein Korpus liefern, zu motivieren. Analog zur Erstellung eines Texts kann die Erstellung eines Korpus aus verschiedenen Blickwinkeln betrachtet werden. Die Erstellung kann als Prozess betrachtet werden oder als fertiges Produkt (Ergebnis) von Erstellungsschritten zu einem bestimmten Zeitpunkt.

Eine ähnliche unterschiedliche Betrachtungsweise kann bei der Untersuchungen zu Modellen der Schreibens und der Textproduktion identifiziert werden. Je nach dem, worauf der Fokus gelegt wird, kann beispielsweise das Schreiben selbst oder das Ergebnis des Schreibvorgangs, der Text, betrachtet werden. Donahue und Lillis (2014) stellen zum Beispiel vier Modelle der `Textproduktion` vor, die verschiedene Perspektiven der Erstellung eines Textes abbilden:

> By "models" of writing and written text production we are referring in a broad sense to the different ways in which the activity of writing and activities around writing are construed. [...] a text-oriented model, a didactic "process" model, a (socio)cognitive model, and a social practices model. (Donahue und Lillis 2014: 55)

Hier wird das Schreiben eines Textes (Textproduktion) aus diesen verschiedenen, auch verschieden zeitlich orientierten Perspektiven betrachtet: Das textorientierte Modell betrachtet das Schreiben als ein fertiges Produkt – der Text selbst ist Gegenstand der Betrachtung. Das didaktische Prozessmodell betrachtet das Schreiben als einen Zeitraum mit Start und Ende. Das sozio-kognitive Modell befasst sich mit dem Lernprozess und dem Verständnis sowie mit dem Schreiben als kognitive und als soziale Aktivität. Letzteres befasst sich mit dem Schreiben als 'practice', also der Tätigkeit des Schreibens.

Interessant für diese Arbeit ist, dass auch hier mit einer zeitlichen Perspektive gearbeitet wird. Wenn der Text als Produkt (text-orientiertes Modell) beschrieben wird, können andere Eigenschaften relevant sein, als wenn der Text als Schreibprozess oder unter kognitiver Perspektive beschrieben wird. Dabei sind zwei der vier

Modelle in dem Rahmen dieser Arbeit mit der Beschreibung von Korpuserstellung vergleichbar, das textorientiere Modell und das Prozessmodell.

Der empirische Untersuchungsgegenstand des textorientierten Modells ist das Produkt, der fertige Text. In linguistischen Analysen werden dann Eigenschaften des Textes auf allen linguistischen Ebenen untersucht (Donahue und Lillis 2014: 56-57). In ähnlicher Weise werden in dieser Arbeit Korpora aus einer produktorientierten Sicht beschrieben und deren Eigenschaften wie bspw. ihre Annotationen dokumentiert.

Das Prozessmodell betrachtet hingegen den Schreibprozess mit allen Phasen des Vorschreibens, des Schreibens und der Revision. Das Prozessmodell eines Textes fokussiert die Aufmerksamkeit auf das Verstehen, wie ein Text erstellt wird, um das Erstellen eines Textes besser zu lernen und zu lehren (Donahue und Lillis 2014: 60). Damit ist nicht der fertige Text Gegenstand der Betrachtung, sonder auch einzelne Schritte im Prozess, die nicht mehr zwangläufig am Produkt selbst erkennbar oder ableitbar sind. Hier können zusätzlich noch Evaluierungen des Prozesses betrachtet werden. So ähnlich verhält es sich mit den Prozessmetadaten, die K. Eckart (2015) beschreibt. Hier geht es unter anderem darum, den Prozess des Annotierens (Parsing) besser zu verstehen und herauszufinden, wie man einen Parser gut auf welche Daten anwenden kann.

Die Metadaten zu a)-c) sind `korpusorientiert`. Da ein Korpus als ein Produkt des Forschungsprozesses und des Forschungsdatenzyklus verstanden werden kann und eine produkt- oder auch ergebnisorientierte Perspektive eingenommen werden soll, werden diese Metadaten in dieser Arbeit `produktorientiert` genannt. Unter Produkt wird hier das Ergebnis verstanden, das aus mehreren Schritten *hervorgebracht* ist. Die Auswahl der zu dokumentierenden Arbeitsschritte wird durch die Zeitpunkte der Veröffentlichung vorgegeben. Auf diese Weise müssten bei einer umfangreicheren Korpusgeschichte nicht alle Schritte mehrfach je veröffentlichter Version chronologisch und exhaustiv aufgelistet werden, sondern nur eine Auswahl bezogen auf das vorliegende Korpus(-produkt). Auch die Bearbeitungsschritte, die nicht direkt Teil des *Hervorgebrachten* sind, wie revidierte Annotationen, werden so ebenfalls nicht erfasst. Die komplette Revisionsgeschichte eines Korpusproduktes ergibt sich dann aus die Menge der Korpusdokumentationen aller Versionen des Korpus.

Nur die Menge an Bearbeitungsschritten aus 1-7 jeweils zu einem gesetzten Zeitpunkt a), b) oder c) werden dokumentiert, die eine permanente Eigenschaft zu diesem Zeitpunkt verantworten (wie Annotationsebene mit Tag C zum Zeitpunkt c).

Die produktorientierten Metadaten beschreiben ein Korpus mit Annotationen und dokumentieren es als eine Art Produkt verschiedener Bearbeitungsschritte, das in Form von definierten festen Versionen veröffentlicht wird. Dies geschieht parallel zum Erstellungsprozess. Die produktorientierten Metadaten abstrahieren von der prozessorientierten Perspektive und beschreiben für Dritte den öffentlichen Stand des Korpus.

Damit richtet sich die hier verwendete Klassifikation von Metadaten nach der in Kapitel 2 erarbeiteten Definition, die Forschungsdaten allgemein als ein gespeichertes und zugängliches Produkt des Forschungsprozesses begreift. Für den Zweck der Wiederverwendung werden nur Eigenschaften von Korpora pro Version benötigt.

4.5 Handlungen durch Metadaten

Auch wenn Metadaten sich immer auf das zu beschreibende Objekt beziehen, ist die Auswahl der Eigenschaften, deren Strukturierung und technische Umsetzung auch von objekt-externen Faktoren bestimmt. Nach Coyle (2005) bestimmt nicht das Objekt selbst die Metadaten, sondern die Anforderungen der Forscherinnen und Forscher, die diese erstellen und nutzen.

Damit sind u.a. Handlungen gemeint, die mithilfe der Metadaten ausgeführt werden können. Dies sind

1. Resource description
2. Information retrieval
3. Management of information resources
4. Documenting ownership and authenticity of digital resources
5. Interoperability

   (Haynes 2004: 15-17).

Metadaten können demnach helfen, eine Ressource zu beschreiben, Informationen zu gewinnen und deren Urheber und Authentizität zu dokumentieren. Mit Hilfe von Metadaten können Informationen gesucht und verwaltet werden. Mit Hilfe solcher multifunktionalen Metadaten kann ein nachhaltiges Teilen von Forschungsdaten gefördert werden (Simons 2014; Simons und Bird 2008). Die Handlungen wiederum werden von verschiedenen Akteurinnen und Akteuren ausgeführt. Wie in Kapitel 3

bereits ausgeführt, ist eine genaue Definition der Akteurinnen und Akteure wichtig, die die Handlungen ausführen, da diese unterschiedliche Kenntnisstände über Korpora besitzen und somit unterschiedliche Handlungen ausführen können. Diejenigen, die ein Korpus erstellen, besitzen einen umfangreichen Kenntnisstand über das Korpus, den sie dann in Form einer Korpusdokumentation (Ressourcenbeschreibung) festhalten. Diese Korpusdokumentation können sie nur für sich selbst erstellen oder anderen zur Verfügung stellen. Im letzteren Fall muss eine Korpusdokumentation auch anderen Forscherinnen und Forscher einen Kenntnisstand über das Korpus vermitteln können (Informationssuche). So sind beispielsweise die Korpusdokumentationen von Anselm[86] und RIDGES[87] online für andere Forscherinnen und Forscher zur Verfügung gestellt, damit diese für sie relevante Informationen über das jeweilige Korpus suchen können. Diese Dokumentationen können die initialen Korpuserstellerinnen und -ersteller ebenfalls in gleicher Weise nutzen.

Wie in Abschnitt 4.3 bereits ausgeführt, gibt es Metadaten, die bestimmte inhaltliche Aspekte einer Ressource beschreiben. So werden für jede dieser Handlungen auch deskriptive, strukturelle oder administrative Metadaten gebraucht. Beispielsweise eignen sich administrative Metadaten, die Informationen über die Korpusersteller nennen und beschreiben, für Punkt 4 und für Punkt 1. Deskriptive Metadaten wie der Titel eines Textes sind relevant für die Punkte 1, 2 und 3.

Um die in Kapitel 3 aufgelisteten Szenarien umsetzen zu können, werden alle hier genannten Handlungen, die mithilfe von Metadaten möglich sind, benötigt. In Abschnitt 4.8 werden die einzelnen Akteurinnen und Akteure, Szenarien und Handlungen miteinander in Beziehung gesetzt.

4.6 Form der Metadaten

Ganz allgemein wird davon ausgegangen, dass Metadaten als Attribut-Werte-Paare, die auch als ein Metadatenset an Eigenschaften mit bestimmten Werten darstellen, verstanden werden (Miller 2011: 4). Abbildung 4.2 zeigt an einem Beispiel, dass Metadaten strukturiert, unstrukturiert oder teilweise strukturiert vorliegen können.

[86] `https://www.linguistics.rub.de/anselm/` (besucht am 22.01.2017).
[87] `korpling.org/ridges` (besucht am 25.01.2017).

Es gibt mehrere Korpora in einer Sammlung. Das Korpus mit dem Namen *RIDGES* enthält Texte über Kräuterkunde. Das Anselm-Korpus ist ein Korpus mit Textzeugen des Passionsdialogs.

Korpus 1	
Name	RIDGES
Text	Kräuterkundetexte

Korpus 2	
Name	Anselm
Text	Passionsdialog

```
<sammlung>
  <korpus n="1">
    <name>Anselm</name>
    <text>Passionsdialog</text>
  </korpus>
  <korpus n="2">
    <name>RIDGES</name>
    <text>Kräuterkundetexte</text>
  </korpus>
</sammlung>
```

Abbildung 4.2: *Beispiel für strukturierte, teilweise strukturierte und unstrukturierte Metadaten. Die Metadaten zweier Korpora können in einem Fließtext unstrukturiert angegeben werden. Zwei Tabellen geben dieselben Metadaten bereits in strukturierter Form an. Strukturiert und maschinenlesbar werden sie in XML angegeben.*

Metadaten können menschenlesbar oder maschinenlesbar vorliegen (Abbildung 4.2). Es gibt freie Strukturen wie Freitext, die bpsw. für Homepages oder technische Berichte genutzt werden. In Abbildung 4.2 wird dies mit dem Freitext illustriert, der über zwei Korpora informiert. Die Angabe der Namen der Korpora ist unterschiedlich gestaltet (*mit dem Namen RIDGES*, *das Anselm-Korpus*). Eine Tabelle kann dieselben Informationen für beide Korpora einheitlich strukturieren (*Name - Wert*). Festere Strukturen können mit Hilfe von Resource Description Framework (RDF) (wie z. B. Bosch und Eckert 2014) oder mit Hilfe von Formaten wie XML (wie z. B. TEI Consortium 2015) umgesetzt werden. Dieselben Metadaten aus dem minimalen Beispiel werden nun XML strukturiert und maschinenlesbar dargestellt (Abbildung 4.2). Weiterhin können Ontologien (Jakus 2013) den Inhalt der Metadaten vorgeben, wie beispielsweise die Einführung von festem Vokabular. Je nach Zweck und Adressatengruppe der Metadaten ist eine bestimmte Form zwin-

gend notwendig.

In dieser Arbeit wird eine menschen- und maschinenlesbare Form der Metadaten mittels eines XML-basierten Formates inklusive verschiedener Validierungsmöglichkeiten umgesetzt, um alle oben beschriebenen Handlungen zu ermöglichen, aber auf die Erarbeitung einer Ontologie verzichtet (Kapitel 7).

4.7 Qualität von Metadaten

Die Erstellung der technischen Berichte, der Homepage oder der Ausfüllung von Metadatenformaten kann ähnlich aufwändig und kostenintensiv wie die Erstellung der eigentlichen Ressource sein.

> Some of the major disadvantages of metadata are cost, unreliability, subjectivity, lack of authentication, and lack of interoperability with respect to syntax, semantics, vocabularies, languages, and underlying models. (Hunter 2003: 319).

Weiterhin sei wenig klar, wie verlässlich und objektiv Metadaten sind und wie ihre Authentifizierung und ihre Interoperabilität gewährleistet werden können.

Der notwendige Umfang der Metadaten ist ebenso wenig allgemein zu definieren. Je nach Objektbezug und Zweck können Metadaten unterschiedlich komplex und umfangreich gefordert sein (vgl. Kapitel 5). Aus vielen dieser Aspekte lassen sich auch Qualitätskriterien für Metadaten ableiten.

Für die Bemessung der Qualität von Metadaten werden verschiedene Ansätze und Vorschläge erarbeitet. So listet beispielsweise die NISO (2007: 61-62) sechs Qualitätsmerkmale für Metadaten auf:

- Metadata Principle 1: Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.
- Metadata Principle 2: Good metadata supports interoperability.
- Metadata Principle 3: Good metadata uses authority control and content standards to describe objects and collocate related objects.
- Metadata Principle 4: Good metadata includes a clear statement of the conditions and terms of use for the digital object.

- Metadata Principle 5: Good metadata supports the long-term curation and preservation of objects in collections.
- Metadata Principle 6: Good metadata records are objects themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

Nach diesen Prinzipien sollen sich Metadaten in einem Standard der Fachgemeinschaft einfügen, Interoperabilität fördern, von einer Autorität kontrolliert werden können und eine klare Aussage über die Nutzungsbedingungen der beschriebenen Ressource machen. Weiterhin sollen Metadaten die langfristige Speicherung von den Ressourcen fördern sowie selbst wiederum Standards für digitale Objekte erfüllen. Diese Kriterien sind sehr allgemein gefasst, womit in Bezug auf die jeweiligen Kontexte Spielraum für die Interpretation oder Bezugnahme auf die konkreten Metadaten gelassen wird. Diese Liste ließe sich erweitern. So stellt die Wiederverwendung von Metadaten für Greenberg et al. (2013) ein weiteres Qualitätsmerkmal dar. Solche Qualitätsmerkmalen können aus unterschiedlichen Perspektiven bewertet werden. Aus der Perspektive der Nutzerinnen und Nutzer ergeben sich drei Gruppen: die Datenerstellerinnen und -ersteller, die ihren Korpora Metadaten zuweisen, die Datennutzerinnen und -nutzer, die die Metadaten als Information über ein Korpus nutzen, und aus Sicht derjenigen, die die Daten zur Verfügung stellen (vgl. Monachini et al. 2004). Eine weitere Perspektive für die Untersuchung von der Qualität der Metadaten ist, wie die gegebene Metadatenstandards von Nutzerinnen und Nutzern konkret angewendet werden oder wie die Metadaten in Anwendungen wie Repositorien eingesetzt werden können. Es gibt Ansätze, die Qualität der Metadaten von zu evaluieren, z.;B. für ISLE Meta Data Initiative (IMDI) (Monachini et al. 2004) und für CMDI (T. Eckart 2016) (vgl. Abschnitt 5.3).

Der hier vorgestellte Ansatz wird dahingehend nur qualitativ geprüft, ob die von NISO (2007) genannten Kriterien berücksichtigt oder erfüllt werden können (Kapitel 7).

4.8 Metadaten für den Zweck der Wiederverwendung

Wie bisher dargestellt, können Metadaten inhaltlich, zeitlich und handlungsorientiert klassifiziert werden. Weiterhin sind die Erstellerinnen und Ersteller sowie die Adressaten der Metadaten wesentlich, da die Erstellerinnen und Ersteller einerseits

die Form der Metadaten und deren Aussagekraft bestimmten und können sich die Auswahl und die Form der Metadaten auch ihren Adressaten richtet. In Verbindung mit ihrem Objektbezug ist es wesentlich, zu welchen Zweck etwas beschrieben wird:

> It may not always be clear how complete a description is needed in a given situation. One extreme would be to use the entire item as the description. (Haynes 2004: 65)

In dieser Arbeit beschreiben Korpusmetadaten historische Korpora zum Zweck der Wiederverwendung. Die Wiederverwendung von Korpora stellt eine komplexe Anforderung dar, die eine extensive und unabhängige Beschreibung erfordert. Daher werden die verschiedenen inhaltlichen Typen von Metadaten nach NISO (2004) als Grundlage und mit einer Erweiterung um die Klasse der technischen Metadaten genutzt. Dabei bezieht sich die inhaltliche Klassifikation auf die zu beschreibenden Korpora und korpusexternen Informationen, die über die eigentliche Ressource hinausgehen. In dieser Arbeit werden so vier Metadatentypen genutzt (vgl. Abschnitt 4.3).

Metadatentyp 1 (deskriptiv)**.** Deskriptive Metadaten, die die enthaltenen Texte und Annotationen eines Korpus beschreiben.
Metadatentyp 2 (strukturell)**.** Strukturelle Metadaten, die das Design des Korpus, der Architektur und seiner Bestandteile beschreiben.
Metadatentyp 3 (adminstrativ)**.** Metadaten, die Urheber und Verantwortliche sowie Ansprechpartner identifizieren und rechtliche Rahmen wie Lizenzen kommunizieren.
Metadatentyp 4 (technisch)**.** Metadaten, die die Realisation eines Korpus in sein(e) Format(e) mit den dafür genutzten Tools und Verfahren erfassen.

Die Erstellung der Metadaten wird typischerweise von Akteurin/Akteur 1 (Initialerstellung) bis Akteurin/Akteur 4 (Initial- und Drittbearbeitung) übernommen. Das Erstellen von Metadaten ist die Voraussetzung für bestimmte Handlungen. Folgende für die Wiederverwendung wesentlichen Handlungen, die mit Hilfe von Metadaten ermöglicht werden, werden in dieser Arbeit nach Haynes (2004) angenommen:

Metadatenhandlung 1 (Deskription)**.** Beschreiben der historischen Korpora.
Metadatenhandlung 2 (Management)**.** Katalogisieren der historischen Korpora.
Metadatenhandlung 3 (Retrieval)**.** Indexieren der historischen Korpora.

Metadatenhandlung 4 (Authentifizierung)**.** Eindeutige Identifizierung und Herkunft, Zuordnung der historischen Korpora.
Metadatenhandlung 5 (Interoperabilität)**.** Übertragen, Konvertieren und Wiederverwenden von Metadaten in anderen Kontexten.

Die Metadaten ermöglichen den Akteurinnen und Akteuren verschiedene Handlungen: Handlung 1 kann beispielsweise als eine Korpusdokumentation an Akteurin/Akteur 3 adressiert und gleichzeitig auch von Akteurin/Akteur 1 (Initialerstellung) und Akteurin/Akteur 2 (Inititalbearbeitung) genutzt werden. Eine Referenzierung über Metadaten als Handlung 4 können alle Akteurinnen und Akteure ausführen, z. B. in Form von Referenzen in einer wissenschaftlichen Publikation. Erst bei einer Menge an Korpora, die mit denselben Metadaten beschrieben werden, werden Handlung 2 und Handlung 3 für hauptsächlich Akteurin/Akteur 3 (Drittbearbeitung) und Akteurin/Akteur 4 (Initial- und Drittbearbeitung) relevant. Diese Akteurinnen und Akteure suchen Korpora in einer Menge von Korpora und benötigen daher Metadaten, die die historischen Korpora katalogisieren und indexieren. Die metadatenbasierten Handlung 1 (Deskription) und Handlung 4 (Authentifizierung) stellen die Voraussetzungen für Szenario 8 (Referenzieren) und Szenario 1 (Analyse).

Die ausführliche Diskussion des Begriffs der Metadaten ist notwendig, weil ihr kompletter inhaltlicher Bezugsrahmen, der noch um eine zeitliche Dimension erweitert wurde, sowie ihre umfangreichen Funktionen und deren Handlungen für alle Wiederverwendungsszenarien relevant sind. Die Abbildung der erarbeiteten Wiederverwendungsszenarien auf die einzelnen metadatenbasierten Handlungen ist ein essentieller Bestandteil.

Wie die Korpora und ihre Wiederverwendungsszenarien, ihre Metadaten, die Handlungen auf Basis der Metadaten und die Akteurinnen und Akteure zusammenspielen, zeigt Abbildung 4.3.

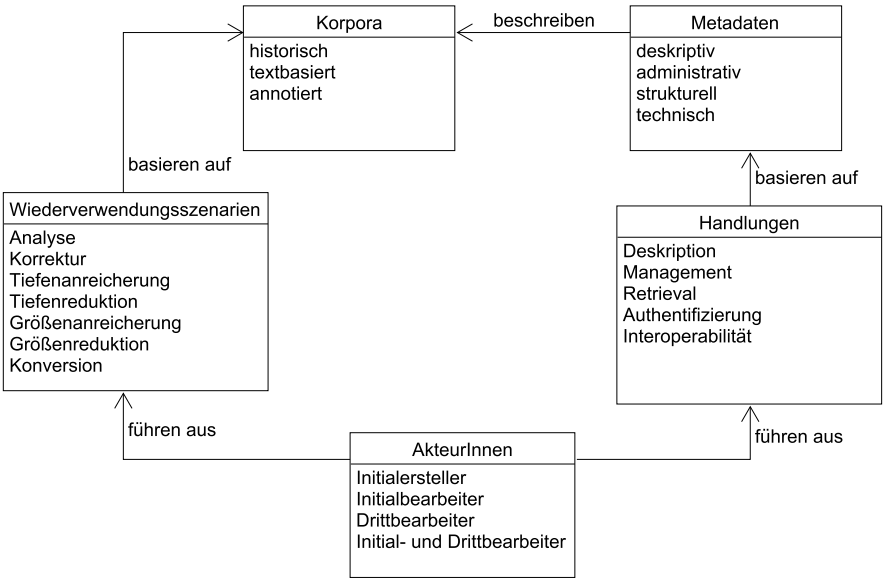

Abbildung 4.3: *Zusammenspiel von Metadaten, Akteurinnen und Akteure, Korpora, Handlungen und Wiederverwendungsszenarien. Wiederverwendungsszenarien basieren auf Korpora,wohingegen Handlungen auf Metadaten basieren. Diese Metadaten beschreiben Korpora. Akteurinnen und Akteure können Korpora wiederverwenden und auf Basis von Korpusmetadaten Handlungen wie die Informationssuche durchführen.*

Akteurinnen und Akteure führen auf Korpusdaten basierende Wiederverwendungsszenarien oder auf Metadaten basierende Handlungen durch. Jedes Wiederverwendungsszenario basiert auf den Korpusdaten selbst. Das Korpus wird z. B. analysiert oder weiter annotiert. Um ein Korpus beispielsweise zu finden, das wiederverwendet werden kann, können Akteurinnen und Akteure auf Basis von Korpusmetadaten nach Korpora suchen und sich über geeignete Korpora informieren (Handlungen). Beispielweise kann so nach Metadaten zum gesuchten Zeitraum der im Korpus enthaltenen Dokumente oder Korpustyp gesucht werden. Dazu müssen die Korpora mit Metadaten beschrieben werden. Wie Akteurinnen und Akteure vor diesem Hintergrund agieren können, zeigt Abbildung 4.4.

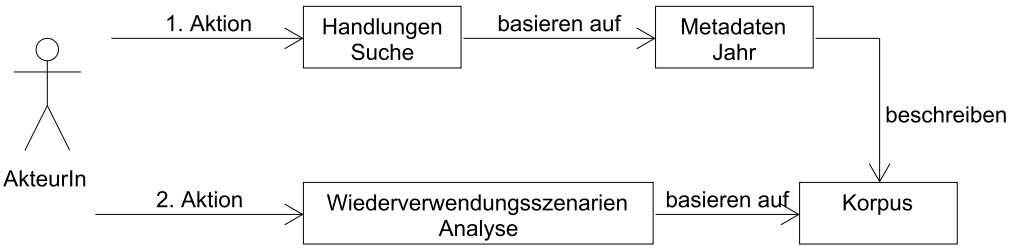

Abbildung 4.4: *Workflow für Akteurinnen und Akteure. Akteurinnen und Akteure können mit Hilfe von Metadaten (Handlung) nach einem Korpus suchen. Das Korpus ist mit diesem gesuchten Metadaten beschrieben. Das so gefundene Korpus kann dann wiederverwendet werden.*

In Abbildung 4.4 führt eine Akteurin oder ein Akteur das Wiederverwendungsszenario Analyse (Szenario 1) an einem Korpus aus (Aktion 2). Eine Voraussetzung, um ein Korpus für eine Analyse wiederzuverwenden ist, dass ein geeignetes Korpus gefunden werden kann. Dies kann eine Akteurin oder ein Akteur mit Handlungen, die auf Metadaten basieren. In diesem Fall sucht eine Akteurin oder ein Akteur nach einem Korpus aus einem bestimmten Jahr (Aktion 1). Diese Handlung basiert auf Metadaten, weil ein Korpus mit diesen Metadaten dokumentiert werden muss und eine Suchsystem auf diesen Metadaten operiert. Es wird nicht im Korpus selbst gesucht, sondern nach bestimmten Metadaten von Korpora. Dazu müssen diese Metadaten mit einer Handlung 1 (Beschreibung) über dieselben oder andere Akteurinnen und Akteure zur Verfügung gestellt worden sein.

Für jedes Szenario ist beispielsweise Handlung 1 (Deskription) essenziell, da diese die Informationen über das Korpus bereitstellt. Um ein Korpus unter einer Menge an Korpora für ein bestimmtes Szenario über eine Metadatensuche auszuwählen, ist Handlung 3 (Retrieval) wichtig.

Die Metadaten erfüllen hier eine Schlüsselaufgabe. Sie können als eine Voraussetzung für die Wiederverwendung von Korpora betrachtet werden. Ein Korpus muss mit Metadaten beschrieben sein, so dass es über diese Metadaten gefunden werden kann und AkeurInnen relevante Informationen über das Korpus erhalten.

In Kapitel 5 wird nun herausgearbeitet, in wie weit bestehende Metadatenstandards für die Dokumentation von historischen Korpora für den Zweck der Wiederverwendung eingesetzt werden können. Dafür werden die für eine Korpusdokumentation relevanten Eigenschaften in Beschreibungskomponenten zusammen gefasst.

5 Metadatenstandards

Metadaten werden in verschiedenen Formen, Funktionen und Umsetzungen in ganz unterschiedlichen Forschungsumgebungen für die Dokumentation von Forschungsdaten angewendet. Neben den Dokumentationen einzelner Projekte in Form von Homepages, Annotationsrichtlinien und Handbüchern gibt es eine Vielzahl an Initiativen und Projekten, die Lösungen für eine allgemeinere oder standardisierte Dokumentation für Forschungsdaten entwickeln. Unabhängig von der Art der Forschungsdaten und den Anforderungen der Fachgemeinschaften an deren Bereitstellung und Archivierung haben alle hier vorgestellten Ansätze das gemeinsames Ziel, durch Metadaten einen Zugang zu Forschungsdaten zu ermöglichen und damit die Nachhaltigkeit von Forschungsdaten zu fördern.

Diese Zielsetzung, die allen Ansätzen mehr oder weniger konkret gemein ist, fassen Jensen et al. (2011) so zusammen:

> [S]tandardisierte Metadaten [sind] eine notwendige Voraussetzung für die Dokumentation und dauerhafte Speicherung von Forschungsdaten. Als Werkzeug fördern sie nachhaltig die Erschließung und Nutzung datenbasierter Forschungsergebnisse. (Jensen et al. 2011: 83)

Es existiert eine enorme Vielzahl an Metadatenstandards (Riley und Becker 2009), die sich zum Teil an dem Forschungsdatenzyklus orientieren. Diese Ansätze werden wiederum in ganz unterschiedlichen Anwendungen eingesetzt, was dazu führen kann, dass Forschung, die auf mehreren dieser Anwendungen fußen will, schwer möglich ist:

> As long as every archive uses its own retrieval interface, has its own metadata structure and tags its content with a proprietary markup schema, research that combines data from many archives becomes intractable. (Küster et al. 2007: Abschnitt 3)

Die verschiedenen Anwendungen besitzen demnach jeweils verschiedenen Suchinterfaces, die auf unterschiedlichen Metadatenschmata basieren. Forscherinnen und

Forscher müssen sich dann in immer wieder neuen Anwendungen zurechtfinden. Die Idee, einem einheitlichen Standard für Archive (und auch vergleichbare Anwendungen) zu entwickeln, verfolgen viele Initiativen und Projekte auf unterschiedliche Weise. Diejenigen darunter, die sich (u.a.) mit Textkorpora oder dokumentorientierten Forschungsdaten befassen, werden daher in den nachfolgenden Abschnitten näher betrachtet.

Die hier vorliegende Arbeit versucht über einen Korpustyp (historische Korpora) und nicht über Anwendungen zu verallgemeinern und für diese Gruppe eine einheitliche Beschreibung zu finden oder ggf. zu entwickeln. Eine Beschränkung auf einen fachspezifischen Kontext (der historischen Korpora) wird in dieser Arbeit nicht angestrebt.

In der vorliegenden Arbeit wurden dafür folgende Anforderungen für eine Korpusdokumentation bzw. an ein Metadatenschema für historische Korpora herausgearbeitet: Ein Metadatenschema muss die Korpusarchitektur von historischen Textkorpora abbilden können. Wie Kapitel 2 gezeigt hat, ist die Beschreibung der Tokenisierung, der Annotationskonzepte und Annotationskategorien wesentlich. Weiterhin muss das Korpus als Produkt seines für ihn speziellen Forschungsdatenzyklus verstanden werden. Wesentlich ist damit die Angabe von Bearbeitungsschritten, die nachvollziehbar zu diesem Produkt geführt haben, und die Dokumentation der Formate und Tools, die zur Erstellung des Korpus genutzt worden sind. Diese Informationen sind für Forscherinnen und Forscher relevant, um Korpora wiederverwenden zu können. Wie Forscherinnen und Forscher Korpora wiederverwenden können, ist mit den verschiedenen Wiederverwendungsszenarien allgemein beschrieben (Kapitel 3).

Für eine Korpusdokumentation, die das ermöglichen soll, werden deskriptive, strukturelle, administrative und technische Metadaten benötigt. Auf solchen Metadaten können dann Handlungen basieren, die eine Voraussetzung für die Wiederverwendung von Korpora sein können (Kapitel 4). So kann ein historisches Korpus über seine Metadaten erst erstellerunabhängig (unter einer Menge an anderen Korpora) herausgesucht und dann über diese Metadaten erschlossen werden, bevor es wiederverwendet wird (Kapitel 1).

Diese Anforderungen werden für einen Vergleich der verschiedenen Metadatenstandards in Beschreibungskomponenten zusammengefasst (Abschnitt 5.1).

Wie weit können also vorhandene Metadatenstandards die herausgearbeiteten Anforderungen einer Korpusdokumentation erfüllen?

Die nachfolgend diskutierten Ansätze stellen unterschiedliche Lösungen für die Dokumentation von dokumentorientierten Forschungsdaten oder Textkorpora bereit.

Sie erarbeiten unterschiedlich stark ausgeprägte oder etablierte Dateninhaltsstandards, Datenwertstandards oder Datenformatstandards (Miller 2011: 13; Gilliland 2008: 3; Zeng und Qin 2016)[88] und werden vor dem Hintergrund der in dieser Arbeit gestellten Anforderungen diskutiert, die sich aus der heterogenen Datenlage und deren Wiederverwendungsszenarien ergeben und durch Metadaten ermöglicht werden sollen (Abbildung 4.3).

Da einige der hier vorgestellten Metadatenframeworks und -formate in mehreren Korpus- und Infrastrukturprojekten eingesetzt werden, werden nur exemplarisch Beispiele für deren Einsatz gegeben (vgl. für einen größeren Überblick zu Umsetzungen von Forschungsdatendokumentationen mit Metadaten auch für andere Arten von Forschungsdaten Miller 2011; Gilliland 2008; Zeng und Qin 2016; Haynes 2004; Baca 2008; Coyle 2005; Jensen et al. 2011; NISO 2004).[89]

5.1 Erfassung von Inhalt, Struktur, Quelle und Bearbeitung der Ressource

Metadaten können deskriptive (inhaltsbezogen), strukturelle, technische und administrative Eigenschaften von Objekten beschreiben (Kapitel 4). Diese Eigenschaften können Eigenschaften des Korpus selbst (`korpuseigen`) oder Eigenschaften von Entitäten sein, die unabhängig vom Korpus existieren (`korpusextern`). Diese Einteilung der Eigenschaften beziehen sich auf die Nähe zum Forschungsprozess (Abbildung 2.2).

Historische Korpora besitzen korpuseigene Eigenschaften wie beispielsweise Namen und Werte der enthaltenen Annotationen. Korpusexterne Eigenschaften werden beispielsweise mit den Erstellern, den genutzten Tools für die Erstellung oder der Art der Veröffentlichung gegeben. Die Angaben, welche historische Texte (Vorlage) in einem Korpus verarbeitet sind, stellen korpusexterne Eigenschaften dar, da ein historischer Text unabhängig von einem Korpus existiert. Dessen Eigenschaften sind hingegen zentral für Teile der Annotationen und damit ist ein Teil auch korpuseigen. So muss ein Geflecht von verschiedenen Objekten, die alle einen un-

[88] Vgl. z. B. Qin und Li (2013) für einen Vergleich von verschiedenen Anwendungen von Metadatenstandards.

[89] Das Projekt DARIAH hat ebenfalls eine umfangreiche Liste von verschiedenen Metadatenformaten zusammengestellt, die einen Überblick über deren Verwendung in den verschiedenen Fächern gibt. `https://wiki.de.dariah.eu/pages/viewpage.action?pageId=38080370#Empfehlungenf%C3%BCrForschungsdaten,ToolsundMetadateninderDARIAH-DEInfrastruktur-Metadatenstandards` (besucht am 21.10.2016).

terschiedlichen Bezug zum Forschungsdatum besitzen, mit Metadaten beschrieben werden (Abbildung 5.1). Diese Objekte werden hier in Beschreibungskomponenten zusammengefasst.

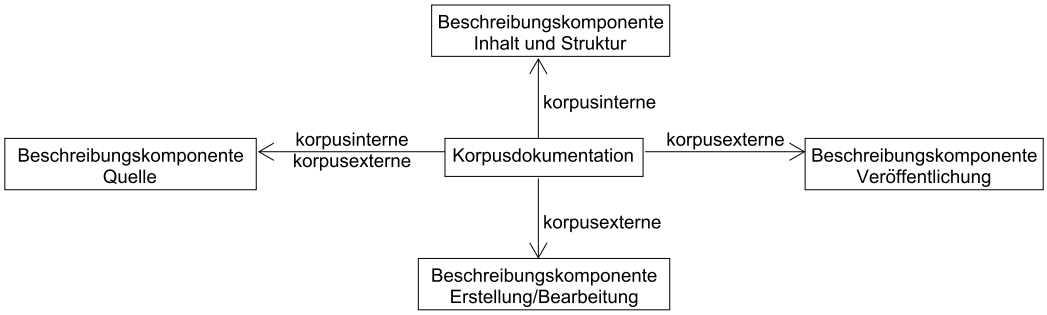

Abbildung 5.1: *Erfassung verschiedener korpuseigener und korpusexterner Beschreibungskomponenten. Dabei können korpuseigene Komponenten wie die Annotationen – Inhalt und Struktur – nicht unabhängig vom Korpus betrachtet werden und sie sind zentral für den jeweiligen Forschungsprozess. Korpusexterne Komponenten wie die Erstellerinnen und Ersteller, die verwendeten Tools und die Veröffentlichungsbedingungen sind unabhängig vom Korpus und weniger ein integraler Bestandteil des Forschungsprozesses. Alle Komponenten sind für eine Korpusdokumentation zum Zweck der Wiederverwendung relevant.*

Abbildung 5.1 zeigt vier Beschreibungskomponenten eines Korpus, welche unterschiedlich stark forschungsbezogen und korpusspezifisch sind. Die Beschreibungskomponente *Quelle* umfasst ganz abstrakt gesehen die Eigenschaften der historischen Vorlagen (Abschnitt 2.7.1). Informationen zur historischen Vorlage sind korpusexterne Eigenschaften, wie auch Metadaten zu deren Autoren oder Editoren. Diese Eigenschaften sind dennoch wesentlich für den Forschungsprozess, da sie einen starken Einfluss auf die Annotationen besitzen können.

Die Beschreibungskomponente *Inhalt und Struktur* fasst die korpuseigenen Eigenschaften wie deskriptive Metadaten zu den enthaltenen Annotationen und deren Beziehung zu einander zusammen (vgl. Abschnitt 2.7.2). Die Komponente *Erstellung/Bearbeitung* umfasst im Wesentlichen korpusexterne Informationen wie z. B. technische Metadaten zu den verwendeten Tools, den Erstellerinnen und Erstellern und korpuseigene Angaben zur Art der Bearbeitung (Abschnitt 2.6). Die Beschreibungskomponente *Veröffentlichung* bezieht sich auf Informationen mit administrativen Metadaten zu Verantwortlichen und Lizenzen.

Wie berücksichtigen die verschiedenen Ansätze diese Beschreibungskomponenten

und die Wiederverwendungsszenarien (Kapitel 3)? In wie weit können sie die Charakteristika von historischen Korpora (Tokenisierung, Annotationskonzepte, Annotationskategorien, Format) überfachlich modellieren und ein Korpus als Produkt eines Forschungsprozesses und als Produkt seiner Bearbeitungsschritte erfassen?

5.2 Dublin Core

Das Dublin Core (DC)-Metadatenschema der DMCI (ISO 2014) besteht aus 15 Elementen, die einen sehr generischen, allgemeinen Ansatz nicht nur zur Beschreibung von einer Vielzahl an Forschungsdaten sondern auch von beispielsweise Tools darstellen können. Dieser Standard wird häufig für Kataloge digitaler Sammlungen oder Bibliotheken verwendet. Welcher Ressourcentyp oder welche Komponenten einer Ressource damit beschrieben werden kann, ist offen, somit können auch historische Textkorpora beschrieben werden.

Tabelle 5.1: *Die 15 DC Elemente mit Erläuterungen.*

Elementname	**Beschreibung**
Contributor	An entity responsible for making contributions to the resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Creator	An entity primarily responsible for making the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Description	An account of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Language	A language of the resource.
Publisher	An entity responsible for making the resource available.
Relation	A related resource.
Rights	Information about rights held in and over the resource.
Source	A related resource from which the described resource is derived.
Subject	The topic of the resource.
Title	A name given to the resource.
Type	The nature or genre of the resource.

Dieser Standard beinhaltet deskriptive, administrative, technische und strukturelle Metadaten. Sie werden als Attribut-Wert-Paare modelliert, die wiederum in vielen Formaten instanziiert werden können.[90] Mit DC werden u.a. der Ersteller, der Herausgeber und der Verteiler, das Format, die Sprache und der Titel der Ressource

[90] `http://dublincore.org/documents/abstract-model/` (besucht am 05.01.2017).

angegeben (vgl. Tabelle 5.1)[91]. Mit Ressource kann beispielsweise die historische Vorlage (Quelle) als ein historisches Korpus insgesamt gemeint sein.

Ein Anwendungsfall des DC-Standards ist die Open Language Archives Community (OLAC)[92]. Diese versteht sich als ein Netzwerk für die Dokumentation und Archivierung von Forschungsdaten und nutzt den DC-Standard für sprachliche Ressourcen:

> OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources. (Bird und Simons 2001: 8)

Ziel ist es also, sprachliche Ressourcen digital zu archivieren und einer breiten Fachgemeinschaft zur Verfügung zu stellen.[93] So ist es anderen Forscherinnen und Forschern mit Hilfe der OLAC möglich, zu fragen: 'Are there any lexical resources for such-and-such a language? (Bird und Simons 2001: 7). Die OLAC nutzt für die Dokumentation von sprachlichen Ressourcen ein XML-basierte Metadatenformat, das sich auf das Dublin Core Metadata Element Set (DCMES) stützt und es zum Teil anpasst sowie erweitert (Bird und Simons 2001). Sie nutzt damit ein sogenanntes *qualified DC*. Der DC-Standard erlaubt den Nutzerinnen und Nutzern die Auswahl der Elemente und ggf. Anpassungen. Beispielsweise ist jedes Element optional und wiederholbar.[94] Damit können die Realisierungen der jeweiligen Metadaten in ihrer Anzahl und Anpassung pro Ressource und pro Ressourcentyp variieren. Nach Salmon-Alt et al. (2006) nutzen viele der Ansätze zur Entwicklung von Metadaten wie OLAC oder IMDI die DC-Metadaten als eine Art Startpunkt, auf dem dann aufgebaut werden kann.

Das DC-Element *Date* ist z. B. mit Verfeinerungen (*Refinements*) angepasst. Mit dem OLAC-*Date* kann genauer zwischen einem Datum für die Erstellung, die Veröffentlichung und die Einreichung der Ressource unterschieden werden. Ein zusätzlich

[91] Elemente der Version 1.1 aus `http://dublincore.org/documents/2012/06/14/dces/` (besucht am 02.12.2016).

[92] `http://www.language-archives.org/` (besucht am 08.09.2016).

[93] Die OLAC hat sich auch darüber hinaus das Ziel gesetzt, nicht nur Korpora sondern auch linguistische Tools beschreiben zu können (Broeder et al. 2002).

[94] Dokumentation des Metadatenformats unter `http://www.language-archives.org/OLAC/metadata.html` (besucht am 08.09.2016).

genutztes Element ist beispielsweise *Rights Holder*, das angibt, welche Person oder Organisation die Rechte an der Ressource besitzt.[95] Über eine Metadatensuche[96] kann nach den jeweiligen angegebenen Metadaten zu ganz unterschiedlichen Typen von sprachlichen Ressourcen gesucht werden. So werden beispielsweise auch Metadaten anderer Ressourcentypen wie dem Analyse et Traitement Informatique de la Langue Française (ATILF) Archive[97] in die OLAC-Metadaten überführt, um ebenfalls über diese Community auffindbar und damit sichtbarer zu werden (Romary und Tucnak 2002).

Ein weiteres Nutzungsszenario des DC-Standards ist das Linguistic Data Consortium (LDC) (Bird und Simons 2003; Cieri und Liberman 2000; Simons und Bird 2008), dessen „primary role was as a repository and distribution point for language resources ".[98] Das DC-Elementeset bzw. die Metadaten sind in dieser Anwendung ebenfalls angepasst. Es werden unterschiedliche Ressourcentypen in Repositorium des LDC wie Transkripte oder komplette Dependenzkorpora mithilfe von DC dokumentiert. Nach diesen Metadaten kann ebenfalls in einer Facetten- und Freittextmetadatensuche gesucht werden.[99]

Da der DC-Standard sehr allgemein und nicht nur für alle Ressourcentypen sondern auch für Tools konzipiert ist, wird er in Anwendungsfällen, wie die hier beschriebenen, angepasst (*qualified* DC). Mit den 15 DC-Elementen allein ist es schwer möglich, ressourcenspezifische Charakteristika für beispielsweise Korpora allgemein zu beschreiben. Mit einem angepassten DCMES werden Bilder, Bücher oder Videos und Korpora sehr allgemein beschrieben. Die DC-Metadaten besitzen weiterhin keine tiefe Strukturierung. Eine Beschreibung der tief strukturierten, korpuseigenen Eigenschaften z. B. von Annotationskategorien -oder Konzepte oder Informationen über die Korpusarchitektur ist so mit DC kaum abbildbar.

Damit besitzt dieser Ansatz grundsätzlich einen geringen Informationsumfang, um alle Beschreibungskomponenten ausreichend zu erfassen. Gerade die Beschreibungskomponenten für den Inhalt und für die Erstellung von historischen Korpora sind komplex und mit einander verbunden. Die Struktur und das Zusammenspiel aller Beschreibungskomponenten historischer Korpora kann nicht allein in den Attribut-Wert-Paaren abgebildet werden. Tief strukturierten Metadaten sind notwendig, um

[95] `http://www.language-archives.org/NOTE/usage.html` (besucht am 27.01.2017).
[96] `http://www.language-archives.org/search` (besucht am 08.09.2016).
[97] `http://www.atilf.fr` (besucht am 27.01.2017).
[98] `https://www.ldc.upenn.edu/about` (besucht am 16.09.2016).
[99] Vgl. z. B. ein Beispiel für einen Katalogeintrag für die Penn Discource Treebank `https://catalog.ldc.upenn.edu/LDC2008T05` (besucht am 05.01.2017).

die wesentlichen Informationen für die Beschreibung/Korpusdokumentation, die AkteurInnen für die verschiedenen Wiederverwendungsszenarien beschreiben zu können.

5.3 ISLE Metadata Initiative und Component MetaData Infrastructure

Die IMDI hat einen Ansatz zur Beschreibung von lexikalischen Ressourcen, linguistischen Ressourcen und Multimedia-Ressourcen entwickelt, der in einem Metadatenset für Katalogbeschreibungen definiert ist (IMDI 2009; Wittenburg et al. 2001):

> While OLAC (Open Language Archives Community) started from a Dublin Core point of view with the goal to create a set that allows for the description of all types of language resources, software tools, and advice, the IMDI (ISLE Metdata Initiative) activities started with a slightly different approach. The focus was primarily on multimedia/multimodal corpora and a more detailed set was worked out that can be used not only for resource discovery but also for exploitation and managing large corpora. (Broeder et al. 2004: 369)

Damit ist dieser Standard spezialisierter als DC und befasst sich allein mit der Dokumentation von unterschiedlichen sprachliche Ressourcen. Die Entwicklung des Metadatensets wurde anhand der Nutzeranforderungen und dem Ressourcentyp entlang entwickelt (Broeder et al. 2002). Dieses Metadatenset beschreibt klassische Attribut-Wert-Paare, die eine Vielzahl an unterschiedlichen Ressourcen beschreiben. „We call the set of metadata elements that describe “published corpora” at the top-level “catalogue” metadata elements for language resources. “(IMDI 2009: 3) Spezifisch für Korpora gesprochener Sprache wurde das Metadatenset für Sessions entwickelt, das strukturelle Metadaten für die Auszeichnungen Einheiten gesprochener Sprache (Session) der Korpora beinhaltet.

> We were guided by the desire to enable not only the resource discovery of major resources such as whole corpora but also be able to find individual resources from within corpora. For instance community members not only want to answer the question “find me all corpora with yaminjung speakers” but also “find me all sessions (recordings) with female yaminjung speakers younger then 60”.(IMDI 2003: 4)

Weiterhin berücksichtigt IMDI die grobe Struktur der Korpora der gesprochenen Sprache, in dem auch einzelne Sprecher und einzelne Aufnahmesessions in den Metadaten aufgeführt werden. Dieser Ansatz geht damit im Vergleich zu DC ein wenig genauer direkt auf die Strukturen des bestimmten Korpustyp ein und erlaubt hierarchische Beziehungen zwischen Metadaten. Weiterhin können die OLAC-Metadaten und die IMDI-Metadaten aufeinander abgebildet werden (Broeder et al. 2004). Ein Anwendungsszenario für das Metadatenset ist das Dokumentation Bedrohter Sprachen (DOBES)-Projekt[100], welches ein Archiv für bedrohte Sprachen entwickelt hat (Wittenburg et al. 2002). Damit sind Audio- und Videoaufnahmen sowie Transkripte die zentralen Forschungsdatentypen. Der Fokus liegt hier auf der Dokumentation gesprochener Sprache in Korpora sowie Korpora selbst.

Aus der IMDI wurde die Component MetaData Infrastructure (CMDI) (ISO 2015) entwickelt, die vorwiegend für linguistische Korpora im Rahmen des CLARIN-Projektes[101] genutzt wird. Damit werden in der CMDI nicht nur Korpora der gesprochenen Sprache sondern auch andere Korpustypen, die in der Linguistik erstellt werden, berücksichtigt. Sie besitzt vergleichbar mit allen hier vorgestellten Ansätzen die Zielstellung, den Zugang, die Wiederverwendung und die Interoperabilität von sprachlichen Ressourcen zu ermöglichen (Broeder et al. 2010: 43).

CLARIN definiert `Metadaten` in Kontrast zu Annotationsdaten („Metadata versus (annotation) data“) wie folgt: „Metadata are believed to be valid and stable for the whole resource it describes.“ (Van Uytvanck et al. 2012: 2) Weiterhin würden Annotationsdaten in Abgrenzung zu Metadaten durch drei Kriterien unterschieden: Annotationsdaten profitieren von einer Darstellung in einem angepassten Tool; sie enthalten elementare linguistische Informationen oder Annotationsinformation (z. B. Tags); sie könnten wortwörtlich in einer akademischen Publikation enthalten sein. Metadaten seien hingegen nicht Teil der Forschungsgegenstands (Van Uytvanck et al. 2012: 2-3).

Diese Unterscheidungskriterien sind aus mehreren Gründen schwierig nachzuvollziehen. Wenn Metadaten über eine Ressource hinweg valide und stabil bleiben sollen, ist dann im Umkehrschluss gemeint, dass Annotationsdaten nicht valide und stabil in den Ressourcen sind? In dieser Auffassung von Metadaten wird nicht thematisiert, dass Korpora in Versionen vorliegen können. Pro Version besitzt ein Korpus jedoch stabile und valide Annotationsdaten. Metadaten können dann einzelne Versionen eines Korpus beschreiben und sind so nur pro Version stabil (vgl. Abschnitt 4.2 und

[100] `http://dobes.mpi.nl/` (besucht am 16.09.2016).
[101] `https://www.clarin.eu/` (besucht am 19.01.2016).

Abschnitt 4.4). Weiterhin ist die Allgemeingültigkeit der Aussage, dass Metadaten nicht zum Untersuchungsgegenstand gehören, in Frage zu stellen. Gerade wenn die Trennung zwischen Daten und Metadaten sehr vage ist, wie auch (Van Uytvanck et al. 2012) feststellen, dann wird dieses Argument damit entkräftet. Auch in Publikationen werden Metadaten, wie z. B. mit welchem Korpus(-teil) welcher Version wie gearbeitet wurde, immer wieder gefordert. Im Übrigen profitieren Metadaten ebenfalls von einer angepassten Darstellung in einem Tool, wie beispielsweise das Virtual Language Observatory (VLO) eine angepasste Darstellung für die CMDI-Metadaten aufweist.[102]

Weiterhin wird zwischen internen und externen Eigenschaften, die mit dem CMDI-Framework beschrieben werden, unterschieden (vergleichbar zu der Unterscheidung in Abschnitt 5.1). Dabei sollten nur externe Eigenschaften direkt mit Metadaten beschrieben werden, interne Eigenschaften sollten separat beschrieben werden und könnten einen Verweis im Metadatenset erhalten (CLARIN-D AP 5 2012: 14-16). Damit werden nicht alle Beschreibungskomponenten in der CMDI aus Abschnitt 5.1 gleich gewichtet. Wie Kapitel 2 in Verbindung mit Kapitel 3 gezeigt hat, sind solche internen (hier mit *korpsueigenen* beschrieben) Eigenschaften der zentrale Baustein, um ein Korpus überfachlich zum Zweck der Wiederverwendung beschreiben zu können. So würde es den Forscherinnen und Forschern wieder überlassen, diese Informationen aufwändig aus den jeweils sehr speziellen und jeweils unterschiedlichen Dokumentationen oder aus dem Korpus selbst herauszulesen.

Die CMDI definiert sich als Framework, mit Hilfe dessen Nutzerinnen und Nutzer eigene, selbst definierte Metadatenformate erstellen können.[103] Dieses Framework stützt sich dabei auf die CLARIN-eigene Metadatendefinition sowie die Organisation von nutzerdefinierten Konzepten und Metadatenformaten durch Datenbanken. Damit wäre dies ein Ansatz für einen Standard zum Datenaustausch. Er besitzt damit keine festen Elementesets und Beschreibungen wie DC oder IMDI.

Mit diesem Framework können Nutzerinnen und Nutzer ein jeweilig unabhängiges XML-basiertes Metadatenformat erstellen, das keinen übergeordneten, festen Guidelines für Elemente und Attribute und deren Struktur folgt oder Einschränkungen bezüglich der funktionalen Klassifikationen von Metadaten macht (vgl. Abschnitt 4.3). Im Vergleich zu den eindeutig definierten DC-Elementen oder der TEI-Guidelines, die für jedes ihrer Elemente eine Beschreibung enthält (vgl. Abschnitt 5.5), kön-

[102] Vgl. die Darstellung der Metadaten von Otfrid: `http://hdl.handle.net/11022/0000-0000-9B20-D` (besucht am 06.01.2017).

[103] `http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml` (besucht am 05.01.2017).

nen Metadatenelemente frei konzipiert und definiert werden. So ist es möglich, dass Elemente mit dem gleichen Namen wie z. B. *Text* unabhängig von einander ganz unterschiedliche Bedeutungen tragen. Diese einzelnen Elementdefinitionen werden in Registraturen organisiert. Somit sollen alle Konzepte der Nutzerinnen und Nutzer gesammelt und werden in eine Data Category Registry (DCR), z. B. in den ISO Catalogue (ISOcat) für CLARIN-D, durch einen Persistent Identifier (PID) referenzierbar eingetragen werden (CLARIN-D AP 5 2012: 4).[104] Ob ein Eintrag eines Elementes von Nutzerinnen und Nutzern öffentlich gemacht wird, hängt davon ab, das CLARIN diesen auch akzeptiert:

> But users can use and create their own components, given that its elements explicitly refer to concepts registered in ISOcat or other trusted registries. If a user wants to include an element which is not yet registered, she would need to register the new concept at least in the so-called ISOcat "user space". The ISOcat process will then decide whether the new category will be integrated in the official part of the registry. CLARIN will be strict and only accept categories that are registered in accepted registries, since otherwise no semantic interoperability can be established. (Broeder et al. 2010: 45)

Um implizite Relationen zwischen einzelnen Elemente explizit zu machen, kann neben der DCR und der Component-Registry noch eine Relational Catalogue genutzt werden. Dort werden nur Beziehungen zwischen Elementen abgebildet (Broeder et al. 2010; Windhouwer 2012).

Gerade bei vielfältig genutzten Konzepten, die als ein Element definiert werden, können Forscherinnen und Forscher ihre eigenen Bezeichnungen für das Element und deren Semantik in der DCR neben anderen hinterlegen (Broeder et al. 2010).

Mit diesem Vorgehen entstehen parallele Entwicklungen in der gleichen Infrastruktur, obwohl „community approved data categories“ ausgewiesen werden, was trotzdem noch nicht in eine standardisierte Datenkategorie oder semantischer Interoperabilität führt (Broeder et al. 2014: 4566-4567). So existieren viele ähnliche aber nicht unbedingt auf einander abbildbare Einträge im DCR (Broeder et al. 2014: 4567), die zumindest ohne für Dritte dokumentierte Beschreibungsebenen vorliegen.

[104] Daneben werden bestehende Standards zusätzlich dazu in diese Infrastruktur importiert. Die CMDI bietet beispielsweise in ihrer Component-Registry drei verschiedene TEI-Profile an, vgl. https://www.clarin.eu/faq/how-can-i-convert-my-tei-headers-cmdi, in denen die jeweiligen Profil-Definitionen und -strukturen in CMDI übernommen werden.

Metadatenelemente können also frei durch Forscherinnen und Forscher definiert und in Komponenten und diese wieder in Profilen zusammengefasst werden. Damit wird in der CMDI kein gemeinsames Metadatenmodell oder Beschreibungsmodell zugrunde gelegt. So können beispielsweise Metadatenkomponenten unterschiedliche Aspekten einer Ressource beschreiben (Broeder et al. 2010: 45). Die Komponenten (und die Profile) können jeweils bedarfsgerecht (neu) zusammengestellt werden. Damit wird die Entwicklung eines übergeordneten, einheitlichen Schemas für bestimmte Ressourcentypen oder Anwendungsfälle nicht unterstützt:

> Durch die Nutzung komponentenbasierter Metadaten wird diese weitgehend feste Trennung von stabilem Schema und variablen Instanzen aufgelöst. Grundgedanke der CMDI ist die Generierung neuer, teils hochspezifischer Schemata durch die Kombination existierender bzw. die Neuerstellung fehlender Komponenten. (T. Eckart 2016: 29)

Bereits entwickelte CMDI-Metadaten sind in der Component-Registry[105] aufgelistet, wobei nicht alle öffentlich zugänglich sind. Forscherinnen und Forscher könnten darin nach vorhandenen Profilen für eine Korpusdokumentation suchen. Ohne eine einheitlich zugewiesene Beschreibungsebene liefert eine Suche nach einem Stichwort *text* für den Korpustyp in der Component-Registry, u.a. zwei *TextCorpus-Profile*, zwei *teiHeader-Profile*, ein *textCorpusProfile*, ein *TextProfile*, ein *AnnotatedCorpus-Profile*, ein *AnnotatedCorpusProfile-DLU* und ein *UnannotatedCorpus-Profile*. Einige dieser Profile sind ausgehend von der angegebenen Beschreibung oder deren Namen für spezielle Projekte entwickelt. Alle Profile, auch die drei *(t/T)extCorpus-Profiles*, enthalten jeweils unterschiedliche Komponenten.

In Bezug auf die Nutzung dieser verschiedenen Profile, müssen Nutzerinnen und Nutzer sich in spezifische Schematisierungen einarbeiten. Automatische Einleseprozesse müssen ebenfalls mit diesen verschiedenen Schematisierungen umgehen können.

> However the interpretation of the TEI files widely depends on the availability of specific schemas. For harvesting metadata from TEI files the specific sub-schema must be known and an extraction has to be done. (CLARIN-D AP 5 2012: 19)

Diese Schwierigkeiten gelten dann ebenfalls für die CMDI-Profile. Darüber hinaus erschweren diese verschiedenen Schematisierungen den Vergleich (oder eine Abbildung) zwischen den CMDI-Profilen/Komponenten oder den von Nutzerinnen und

[105] `https://catalog.clarin.eu/ds/ComponentRegistry` (besucht am 11.12.2016)

Nutzern selbst erstellten Elementen innerhalb der CMDI und zwischen anderen externen Standards.

Daher werden hier nur drei öffentliche Profile beispielhaft betrachtet und mit den gestellten Anforderungen abgeglichen. Diese TextCorpus-Profile beziehen sich auf die Ebene des Korpus als Sammlung (top-level). *textCorpus*, *TextCorpus'*, *TextCorpus"* oder *Corpus* sind jeweils eigenständige CMDI-Profile mit unterschiedlichen Komponenten. Ein TextCorpusProfile ist ein "Corpus profile for all text collections with or without annotations„[106], ein anderes TextCorpusprofile ist "A CMDI profile for text (i.e. written) corpus resources.„[107] und ein drittes "Clarin-NL profiles„[108].

Diese Profile nutzen unterschiedliche Elemente, die in unterschiedlichen Strukturen (Komponenten) mit einander verbunden sind und die wiederum mit KonzeptLinks ausgestattet werden können, aber nicht müssen.[109] Viele dieser Konzepte basieren auf Einträgen in der ISOcat-Registry. Kann der Link aufgelöst werden, so erhält man die Erklärung zu einem Konzept wie *Text*. Wenn man nach Konzepten wie *Text* in der dazugehörigen ConceptRegistry sucht, erhält man viele Einträge und unterschiedliche Bedeutungen.[110] Für *Text* gibt es 205 Treffer, die auf ganz unterschiedliche Arten das Konzept *Text* einbinden. Vor dem Hintergrund von Abschnitt 2.7, ist dies nicht überraschend. Die Orientierung über mehrere Profile hinweg, z. B. welches Konzept mit welchem Element in welcher Komponenten und in welchen Profilen genutzt wird, bleibt den Anwenderinnen und Anwendern überlassen. Ein Vergleich über mehrere Profile müsste alle Konzepte aller Elemente, aller Gruppierungen der Elemente zu unterschiedlichen Komponenten und diese zu unterschiedlichen Profilen berücksichtigen.

So erlaubt das CMDI-Framework unendliche viele Einträge von Elementen in der DCR (mit ggf. dem gleichem Elementnamen), die wiederum mit unterschiedliche Relationslinks zu anderen frei definierten Elementen ausgestattet werden können. Durch die Gruppierung der Elemente in Komponenten und Profilen wächst die Vielfalt weiter an. Diese freie Ausgestaltung des Struktur und des Inhalts von CMDI-Metadaten ist in keine einheitliche Beschreibungs- oder Instanziierungsebene für sprachliche Ressourcen (und deren Dokumente und Annotationen) zusammengefasst. Korpora desselben Korpustyps können mit unterschiedlich strukturierten und

[106] `clarin.eu:cr1:p_1386164908461` (besucht am 22.12.2016).
[107] `IDclarin.eu:cr1:p_1290431694580` (besucht am 22.12.2016).
[108] Die Abkürzung NL steht Niederlande. Gemeint ist ein CLARIN-Profil des niederländischen Projektteils. `clarin.eu:cr1:p_1271859438164` (besucht am 22.12.2016).
[109] Ein Konzeptlink wird als eine PID in einem Attribut des entsprechenden Elements angegeben.
[110] `https://openskos.meertens.knaw.nl/ccr/browser/index.php?key=text&termsOr=true&matchTermsExact=true&facet0=ALL&facet3=ALL&facet4=meertens` (besucht am 21.10.2016).

inhaltlich variierenden Metadatensets beschrieben werden. Zum Beispiel werden das Korpus GerManC[111] und das Korpus des DTA[112] mit demselben Schema beschrieben, das Korpus Otfrid[113] mit einem anderen. Nach Kapitel 2 sind diese drei Korpora alle vom gleichem Typ, historisches Textkorpus. T. Eckart (2016: 165) stellt ebenfalls eine „unkontrollierte Nutzung unterschiedlichster Designentscheidungen für identische Dokumentationsinteressen“ fest. Nicht klar nachvollzogen werden kann, ob die jeweiligen Profile tatsächlich dieselben Dokumentationsinteressen besitzen oder jeweils aus unterschiedlichen Szenarien heraus erstellt wurden.

Die bereits vorhandenen Profile für Textkorpora (soweit zugänglich) bilden nur einen Teil der gewünschten Beschreibungskomponenten ab. Dies ist nicht weiter überraschend, da die CMDI-Metadaten per Definition wenige korpuseigenen Informationen tragen (sollen).

Die Erstellung eines eigenen CMDI-Profils ist ohne eine eigene abstrakte Beschreibungsebene schwer möglich. Die CMDI besitzt kein eigenes Metamodell. Jede CMDI-Nutzerin und jeder CMDI-Nutzer verwendet implizit oder explizit eigene Modelle, so dass auf unter Umständen konkurrierenden, ähnliche oder nicht auf einander abbildbaren Bedeutungen verwendet werden. Die Zusammenstellung von Elementen zu Komponenten und deren Zusammenstellungen zu Profilen führt dann ebenfalls zu nebeneinander stehenden Ansätzen zur Dokumentation von sprachlichen Ressourcen oder Teilen davon. Einzelne CMDI-Profile geben damit unabhängig von einander eine Schematisierung vor. So könnte dasselbe Element mit demselben Konzeptlink in unterschiedlichen Strukturen und Beschreibungsebenen eingebettet werden (vgl. Abbildung 5.2).

[111] http://hdl.handle.net/11372/LRT-882
[112] http://hdl.handle.net/11372/LRT-172
[113] http://hdl.handle.net/11022/0000-0000-9B20-D

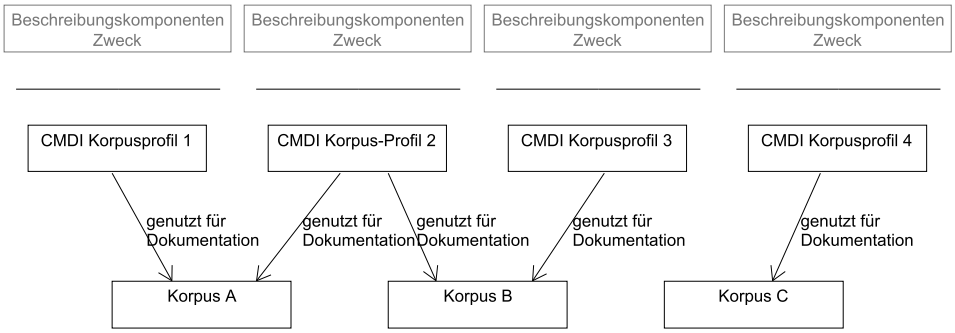

Abbildung 5.2: *CMDI-Profile basieren jeweils auf Beschreibungskomponenten eines Korpus für einen bestimmten Zweck, die jeder Metadatenerstellerinnen und -ersteller für seinen eigenen Ressourcentyp festlegt. Ein Profil kann mehrere Korpora beschreiben. Ein Korpus kann durch ein oder mehrere Profile beschrieben werden.*

Ein Profil kann als korpusspezifisch oder projektspezifisch verstanden werden, wenn es genau ein Korpus beschreibt (oder beschreiben kann). Ein Korpus kann auch durch mehrere dieser Profile dann u.U. jeweils anders beschrieben werden. Jedes CMDI-Profil verwendet möglicherweise andere Beschreibungsebenen für unterschiedliche Zwecke. Diese Profile unterstützen aber nur einen Teil der geforderten Beschreibungskomponenten. Damit stellt CMDI genau eine Infrastruktur für die Erstellung von Metadaten dar, die weder eigene einheitliche Modelle oder Beschreibungsebenen den Nutzerinnen und Nutzern zur Verfügung stellt noch eigene versionierte Guidelines für die Beschreibung von Elementen festlegt. Nutzerinnen und Nutzer müssen sich Korpora dann über diese unterschiedlichen Profile und die unterschiedlichsten Konzepten sowie Beschreibungsebenen erschließen. Metadatenerstellerinnen und -ersteller müssen eigene Modelle und Beschreibungskomponenten erarbeiten, die sie dann mit darauf zugeschnitten Elementen, Komponenten und Profilen realisieren können. So kann dieser Ansatz den gestellten Anforderungen nicht gerecht werden.

5.4 Metadata Encoding and Transmission Standard

Der Metadata Encoding and Transmission Standard (METS) ist für die Dokumentation der Struktur von Objekten einer digitalen Bibliothek entwickelt worden, der einen starken Fokus auf die verschiedenen Beschreibungsebenen eines Dokumentes besitzen. Damit ist METS wie DC ein Standard, der Datenstrukturen

beschreibt. Bei den Objekten handelt es sich um Bilder oder Bilder-Text-Objekte wie z. B. die digitale Sammlung historischer Zeitung der Universität Heidelberg [114]:

> The Metadata Encoding and Transmission Standard (METS) is a data encoding and transmission specification, expressed in XML, that provides the means to convey the metadata necessary for both the management of digital objects within a repository and the exchange of such objects between repositories (or between repositories and their users). (METS Primer 2010: 15)

Dieser Standard ist eine XML-basierte Spezifikation für die Kodierung und Übertragung von Daten. Dieser Standard soll eine Übertragung von Metadaten zwischen Repositorien und zwischen Repositorien und Nutzerinnen und Nutzern ermöglichen. Der METS wird in sieben Abschnitte unterteilt: (1) METS-Header mit Metadaten über das METS-Dokument selbst, (2) deskriptive Metadaten zur Erschließung des Dokuments, die mit DC-Metadaten angegebenen werden, (3) administrative Metadaten zu unter anderem der Urheberschaft der Daten, (4) Liste aller Dateien, des zu beschreibenden Objektes, (5) eine Strukturbeschreibung, die den inneren Aufbau des digitalen Objektes beschreibt, (6) eine Linksammlung für die Strukturbeschreibung und (7) eine Linksammlung zu Anwendungen, die beispielsweise Objekte anzeigen lassen können (Jensen et al. 2011; METS Primer 2010; NISO 2004):

METS kann in verschiedenen Profilen spezifisch angepasst und mit verschiedenen anderen Standards wie DC verknüpft werden (Jensen et al. 2011) und wird so in den unterschiedlichsten Bibliotheken [115] genutzt.

METS weist klare Beschreibungskomponenten (dokumentorientiert) und einen klaren Zweck auf. Verschiedene funktionale Typen von Metadaten werden in diesem Standard integriert. Die Lösung, die METS vorschlägt, besitzt einen anderen Objektbezug als in Abschnitt 5.1 definiert. Es wird sehr umfangreich die Beschreibungskomponente *Quelle* für die Dokumentation in Bibliotheksanwendungen berücksichtigt. Hier stehen bibliographische Metadaten zu Objekten einer digitalen Bibliothek im Vordergrund. Dabei sind häufig die verschiedenen Ebenen einer Publikation und die Definitionen von Text oder Dokument, wie sie beispielsweise die FRBR vorschlägt, zentral (vgl. Abschnitt 2.7.1).[116]

[114] `http://digi.ub.uni-heidelberg.de/diglitData6/mets/allgemeine_politische_nachrichten.xml` (besucht am 11.12.2016).

[115] `http://www.loc.gov/standards/mets/mets-registered-profiles.html` (besucht am 16.09.2016).

[116] Vergleichbare Ansätze für Datenstrukturstandard dazu geben Metadata Object Description Schema (MODS) und MAchine-Readable Cataloging (MARC), die ebenfalls ein bi-

Eine weitere Komponente, die damit eng verknüpft ist, stellt die verschiedene Sicht auf das digitale Surrogat einer (historischen) Vorlage dar. Eine historische Vorlage (z. B. ein Buch) kann in unterschiedlichen digitalen Surrogaten und Surrogatausschnitten wie Text der Bild vorliegen. Historische Korpora beinhalten Text(e) oder ein oder mehrere digitale Repräsentationen von Texten. Korpora stellen in so fern auch eine Art digitales Surrogat einer historischen Vorlage dar (Abschnitt 2.7). Da die Annotationen in Textkorpora diese Repräsentation realisieren, sind sie ein wesentlicher Bestandteil, der in METS nicht berücksichtigt wird und aber ebenfalls mit Metadaten beschrieben werden muss. Die Komponenten, die ebenfalls relevant für historische Korpora sind, sind die Erstellung und der Inhalt (in Bezug auf Annotationen). Diese sind weniger ausführlich in den Ansätzen von Metadatenstandards wie METS für Bibliotheken abgebildet. Damit kann dieser Standard auch nicht alle Anforderungen bedienen.

5.5 Text Encoding Initiative

Die TEI ist eine überfachliche Initiative, die Guidelines für die Digitalisierung von Dokumenten entwickelt und u.a. in den Literaturwissenschaften, den Geschichtswissenschaften, den Sozialwissenschaften und der Linguistik genutzt wird.[117] Die TEI ist primär ein Standard für das Markup von Dokumenten als ein alleiniger Standard für Metadaten (Zeng und Qin 2016: 16).

In diesen Guidelines werden ganz unterschiedliche Konzepte mit Bezug auf die Digitalisierung und Auszeichnung von Dokumenten definiert, womit diese Guidelines als eine Art Dateninhaltsstandard gesehen werden können.[118] Neben den Guidelines existiert ein gleichnamiges XML-basiertes Format und eine Modellumgebung mit der TEI-Spezifikation One Document does it all (ODD), durch welche das Format entweder als ein originäres Subset der TEI oder als spezifische Erweiterung modelliert und ein dazu passendes Validierungsverfahren in Form von Schemata erzeugt werden kann (Burnard und Rahtz 2004). Damit existiert zu diesem Dateninhaltsstandard auch ein Standard für den Datenaustausch.

Durch die Zusammenarbeit einer breiten Fachgemeinschaft werden die TEI-Guidelines

bliographisches Elementeset für eine Vielzahl an solchen Objekten entwickelt. Vgl. für MODS `http://www.loc.gov/standards/mods/mods-overview.html` (besucht am 11.12.2016) und für MARC NISO (2004).

[117] `http://www.tei-c.org/` (besucht am 20.01.2017).

[118] Es gibt darüber hinaus die mit der TEI assoziierte Music Encoding Initiative (MEI), die sich auf musikalische Dokumente wie Notenblätter spezialisiert (`http://music-encoding.org` (besucht am 21.10.2016).

stetig weiterentwickelt und so fachspezifische Anforderungen in Form von neuen Elementen und Modulen inkorporiert (Burnard 2013: 11,34), ohne dabei eine zu enge und zu fachspezifische Umsetzung zu erreichen. Zum Beispiel werden aus der Fachgemeinschaft heraus spezialisierte TEI-Module wie z. B. `epiDoc` (Bodard 2010) entwickelt, um fachlichen Anforderungen gerecht zu werden. Damit besitzt sie eine hohe Akzeptanz über verschiedene Fächer hinweg. Durch diese Erweiterungen gibt es immer mehr Korpustypen, die mit der TEI digitalisiert und ausgezeichnet werden können.

> One of the building blocks of the TEI's success among various scholars is the fact that it does not define a normative standard but rather guidelines. These recommendations try to not constrain the user to a single way of encoding but leave a large amount of personal freedom (and responsibility) to the user, while other annotation formats try to be as strict as possible to reflect a certain annotation model and theory. (Stührenberg 2012: 9)

Die TEI-Guidelines sind primär eine Art Annotationsrichtlinie, die flexibel für die Ausweisung von digitalen Ressourcen genutzt werden kann, wie es Abschnitt 2.3.2 exemplarisch für die Auszeichnung von Autoren zeigt. Mit ihren mehr als 500 Elementen, deren vielfältigen Attributen und verschiedenen Modulen, die Elemente und Attribute gruppieren, ist die TEI flexibel und umfangreich. Typische Merkmale von Dokumenten, die mit Hilfe der TEI ausgewiesen werden, sind graphische Eigenschaften des Texte, wie Zeilen- und Seitenumbrücke, Markup und Textgestaltung.

Ein mögliches Ziel, das mit der TEI umgesetzt werden kann, wäre die digitale Repräsentation eines historischen Texts (digitales Surrogat) mit einer möglichst diplomatischen, gemeint ist eine nahe an der historischen Vorlage orientierte Darstellung des Inhalts mit kritischen, fachbezogenen Anmerkungen zu Interpretationen und Hintergründen (Baillot und Seifert 2013). Viele Projekte, die sich auf diese Richtlinie stützen, nutzen ein angepasstes Schema, eine Art Subset aus den kompletten TEI-Guidelines. Eine solche spezifische Schematisierung der TEI ist mit Hilfe der TEI-ODD möglich (Burnard und Rahtz 2004). Durch die ODD besitzt die TEI eine Modellierungssprache im Vergleich zu allen anderen Ansätzen, mit der eine Anpassung eines Schemas auf einer abstrakten Ebene möglich ist. [119]

Da sich die TEI stark auf die Quelle (hier historische Vorlage) bezieht, gibt es in

[119] Im Kapitel 7 wird die Funktionsweise näher erklärt.

jedem TEI-konformen Dokument die Möglichkeit, umfangreiche Angaben zur Veröffentlichung und Überlieferung der Vorlage anzugeben.

Die Angabe von umfangreichen Informationen (Metadaten) zum Digitalisat und seinem wissenschaftlichen Kontext ist ein zentraler Punkt für die Erstellung einer TEI-konformen Datei. Für die Auszeichnung von Metadaten besitzt die TEI ein eigenes Modul `header`, das mit dem Element `<teiHeader>` in der TEI-konformen Datei eingesetzt wird:

> <teiHeader> (TEI-Header (elektronische Titelseite)) Beschreibungen und Erklärungen, die eine elektronische Titelseite für ein TEI-konformes Dokument ergeben. [120]

Der Zweck des `teiHeader` ist also die Beschreibung eines TEI-konformen Dokumentes und seines Inhalts. Die zugrundeliegende Struktur bildet eine TEI-XML-Datei mit einem digitalisierten und annotierten Text – einem Dokument – sowie dessen Metadaten, die im `teiHeader` angeben werden. Damit beziehen sich die Metadaten, die in dem `teiHeader` angegeben werden, auf das Dokument – den Text – in derselben TEI-konformen Datei (Abbildung 5.3).

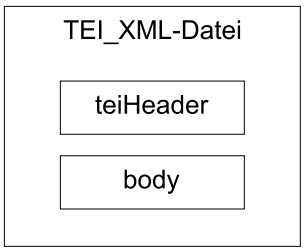

Abbildung 5.3: *TEI-Dokument mit `teiHeader` und Text.*

Der `teiHeader`[121] kann mit unterschiedlichen deskriptiven, administrativen und technischen Metadaten in Form von fünf Komponenten ausgestattet werden:

> **fileDesc** (Dateibeschreibung) enthält die detaillierte bibliografische Beschreibung einer elektronischen Datei.

[120] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-teiHeader.html` (besucht am 15.09.2016).

[121] Hier stelle ich nur die für diese Arbeit relevanten Teile vor. So wird die Komponente `<xenoData>` nicht weiter beschrieben, vgl. `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-xenoData.html` (besucht am 02.12.2016).

encodingDesc (Beschreibung der Kodierung) dokumentiert das Verhältnis zwischen dem elektronischen Text und seiner Quelle oder den Quellen, von der oder von denen er abstammt.

profileDesc (Beschreibung des Textprofils) enthält eine genaue Beschreibung der nicht bibliografischen Kennzeichnungen des Texts, besonders der verwendeten Sprachen und Subsprachen, der Entstehungsbedingungen, der Teilnehmer und ihres Umfelds.

sourceDesc (source description) beschreibt den (die) Quelltext(e), von dem (denen) der elektronische Text abstammt oder erzeugt wurde.

revisionDesc (Beschreibung der Textrevision) enthält alle Revisionsschritte, die an der Datei vorgenommen wurden.[122]

Diese Elemente wiederum können mit weiteren Elementgruppen oder Modulen wie z. B. `msDescription`[123] für die Beschreibung der Publikationshistorie einer Handschrift erweitert werden.

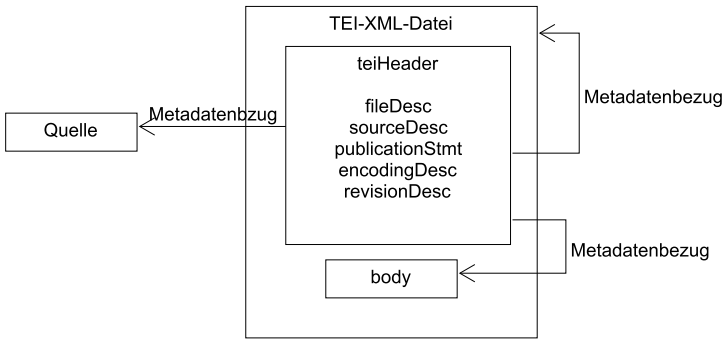

Abbildung 5.4: *Beschreibungskomponenten im* `teiHeader`*. Der Teil* `sourcedesc` *bezieht sich auf die Beschreibungskomponente* Quelle*. Der Teil* `fileDesc` *bezieht sich dabei auf die TEI-XML-Datei. Der Teil* `encodingDesc` *beschreibt das Verhältnis zwischen Datei und Quelle. Angaben über die Veröffentlichung der TEI-XML-Datei stehen im Teil revisionDesc.*

Daraus resultieren mehrere Komponenten, die für die Beschreibung eines Dokumentes relevant sind. Nach Romary (2013) sind das drei Hauptkomponenten: die

[122] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/HD.html#HD11` (besucht am 21.10.2016).

[123] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-msDesc.html` (besucht am 08.01.2017).

Datenerstellung, das Datenmanagement und deren Veröffentlichung, die mit Hilfe des `teiHeader` in einem TEI-Dokumentes adressiert werden (Abbildung 5.4).

Ein Beispiel für ein Repositorium, das TEI-konforme Dokumente archiviert und dokumentiert, ist TEXTGRID von DARIAH (Hedges et al. 2013):

> [...] TextGridRep enables researchers to publish and share their data in a way that supports long-term availability, interoperability, and reusability. (Hedges et al. 2013: Abschnitt 2, Paragraph 9)

Das Textgridrepositorium hat sich neben der Archivierung von Bildmaterialien auf TEI-basierte Korpora aus verschiedenen geisteswissenschaftlichen Fächern spezialisiert. Textgrid nutzt dabei für die Dokumentation der im Repositorium enthaltenen Daten nicht die Angaben in den Headern selbst, sondern ein eigenes XML-basiertes Metadatenformat:

> We are well aware that we cannot force the TextGrid metadata structure, our markup schemata etc. onto existing archives, of course. But we will start with a minimal, DublinCore-inspired subset of metadata that presumably can be mapped to the metadata structure of the relevant archives. If the metadata subset that can be mapped between TextGrid and the external provider is larger, more complex retrieval queries from within TextGrid can be handled. (Küster et al. 2007: Abschnitt 3)

Über deskriptive (Titel, Autor, Sprache, Jahr), administrative (Lizenzträger) und technische Metadaten (Format) kann nach TEI-XML-Dokumenten in dem Repositorium gesucht werden. Die TEI selbst wird nicht zur Erstellung der Metadaten im Repositorium genutzt.

Damit wird die TEI selbst wenig als Metadatenstandard genutzt. Die TEI besitzt einen konkreten Fokus, das Dokument. Damit bezieht sich ein `teiHeader` typischerweise auf Dokumenteigenschaften und Dateieigenschaften und weniger auf Korpuseigenschaften. Diese Dokumente können aber Teil eines Korpus sein (vgl. Kapitel 2). Im `teiHeader` werden weiterhin kaum Eigenschaften von Annotationen in einem Korpus berücksichtigt. Die Annotationen in einem TEI-Dokument werden nicht zwangsläufig im `teiHeader` beschrieben und es gibt hierfür auch keine ausgewiesene Stelle, da sie bereits in den TEI-Guidelines beschrieben sind. Innerhalb der TEI-Welt ist eine derartig umfangreiche Dokumentation durch Metadaten nicht derart notwendig, da TEI-konforme Dateien der TEI-Guidelines bereits folgen.

Einen neuen Ansatz, diese Lücke zu schließen, verfolgen Bański et al. (2016) mit der Integration von einem Standoff-Module in die TEI-Architektur.

Dennoch liefert TEI mit dem `teiHeader` eine umfangreiche Lösung für die verschiedenen funktionalen Klassen von Metadaten, die auch in dieser Arbeit benötigt werden. Weiterhin stellt die TEI indirekt einen Dateninhaltsstandard für die Metadaten eines digitalen Dokuments, die im `teiHeader` angegeben werden können. Dabei können die Perspektiven auf digitale Surrogate sowie eine ggf. komplexe Überleiferungsgeschichte einer Quelle berücksichtigt werde.

Für die Verwendung von TEI ist eine abstraktere Modellierung notwendig, um die Definition von Korpus, seiner Dokumente und Annotationen inklusive seiner Wiederverwendungsszenarien in der TEI-Welt abzubilden. Nur so können auch die umfangreichen Möglichkeiten, Metadaten nach den TEI-Guidelines (Inhaltsstandard) auszuwählen, die die gewünschten inhaltlichen, zeitlichen und funktionalen Aspekte erfüllen, und danach in TEI-XML (Datenaustausch) umzusetzen.

Es gibt meiner Kenntnis nach wenige Ansätze, die die TEI-eigene Lösung für Metadaten auch außerhalb der TEI-Welt nutzen: Mit dem `teiHeader` besitzt die TEI eine strukturierte Umsetzung für die dokumenteigenen Metadaten.

5.6 Diskussion

Die vorgestellten Metadatenstandards DC und IMDI beziehen sich mit einer Art Top-Level-Ebene auf digitale Objekte und können die komplexe Architektur von Korpora mit den verschiedenen Beschreibungskomponenten mit den sehr allgemeinen Elementesets nicht umfassend genug erfassen. Die CMDI liefert als eher ein Datenübertragungsstandard keine strukturell oder inhaltlich einheitlichen Modelle, auf die für die Dokumentation von historischen Korpora aufgebaut werden kann. Die vorhandenen Profile folgen der funktionalen Definition von Metadaten, dass wenige oder keine korpusinternen Angaben zu Annotationskonzepten oder -kategorien in das Schema integriert werden sollen. Damit enthalten diese Profile nicht zu allen geforderten Beschreibungskomponenten Metadaten.

Der METS ähnlich wie die TEI konzentrieren sich stark auf das digitale Dokument. Der METS ist auf Bilder und Bild-Text-Objekte spezialisiert, die nur wenige Eigenschaften mit Korpora teilen. Die TEI ist als eher Dateninhaltsstandard für Dokumente konzipiert und besitzt umfangreiche Guidelines, die komplexere und spezifische Angaben für Dokumente ermöglicht. Sie erfasst mit ihrem Ansatz hingegen nur einen Bestandteil eines Korpus – das Dokument. Ihr Metadatenmodell bezieht

sich stark auf das TEI-konforme Dokument.

Allen Standards fehlt ein großer Teil der Metadaten der Beschreibungskomponenten für den Inhalt und die Struktur, die für die Beschreibung der Korpusarchitektur (Tokenisierung, Annotationskonzepte, Annotationskategorien). Einige Standards enthalten Metadaten für die Beschreibungskomponenten für Erstellung, diese sind dann aber auf die Korpusebene als *top-level* Beschreibung bezogen und enthalten daher nicht tief strukturierte Metadaten zum Inhalt.

Viel hängt zusätzlich davon ab, wie die kritischen Konzepte wie `Text` und `Dokument` erfasst werden können. Korpora bzw. die digitalisierten Texte in Korpora können ebenfalls ein digitales Surrogat einer Quelle (historischen Vorlage) darstellen. Diese Quellen können demnach durch die FRBR erfasst und mit bibliographischen Metadaten versehen werden, wie sie auch DC oder METS vorsehen. Der TEI-Ansatz integriert noch umfangreicher die Beziehung zwischen Quelle und digitalem Surrogat, welche für die Texte in einem Korpus ebenfalls relevant ist.[124]

Ein historisches Korpus besitzt mit der tatsächlichen Umsetzung der verschiedenen Vorlagen noch ein weiteren Aspekt – den der Korpusarchitektur (mit Tokenisierung, Annotationskategorien und -konzepte). Diese sind ebenfalls wichtig für die Korpusdokumentation. Dabei kann auch für historische Korpora keine klare, einheitliche Definition für `Text` gefunden werden (vgl. Kapitel 2). Das TEI-Dokument und seine Beschreibungskomponenten reichen für den in dieser Arbeit aufgezeigten Anforderungen nicht ganz aus.

Es fehlt also ein Konzept, wie das Geflecht aus Quellen und ihren Metadaten, aus digitalen Surrogaten und ihren Metadaten sowie Annotationen und ihren Metadaten in einer Korpusdokumentation organisiert werden kann. Ein solches Konzept kann in Form eines Metamodells erstellt werden, das die notwendige abstrakte Beschreibungsebene besitzt. Dieses Metamodell kann beschreiben, welche Metadaten für historische textbasierte Korpusdaten für den Zweck der Wiederverwendung notwendig sind (Abschnitt 4.2). Dieses Konzept kann nicht aus den verschiedenen Projekt- oder Infrastrukturen selbst emergieren[125], sondern muss modelliert werden.

[124] Vgl. zu einer Integration dieser verschiedenen Ansätze, wie z. B. dokumentoriente Ansätzen für digitale Bibliotheken Romary (2012).

[125] Wie van Zundert (2012) bemerkt, sind das Kodieren und Modellieren zentrale Bausteine für die Forschung, nicht die Projekt- oder Infrastrukturen.

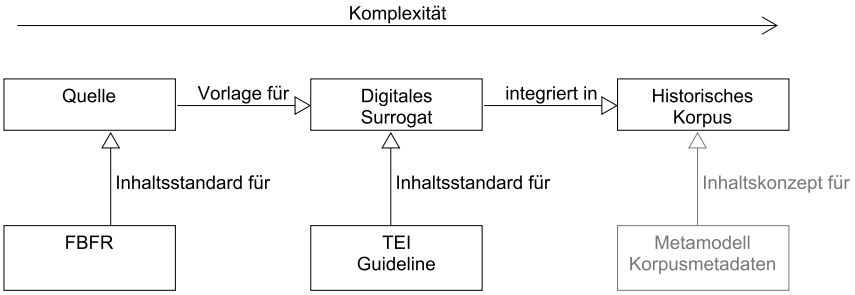

Abbildung 5.5: *Beziehung zwischen Quelle, digitalem Surrogat und historischem Korpus. Die Metadaten einer historischen Quelle sind mit dem Ansatz der FRBR beschreibbar. Die TEI beschreibt mit ihrem Ansatz ein digitales Surrogat (Dokument) und bezieht dabei auch die Beschreibungen der Quelle mit ein. Ein historisches Korpus wiederum integriert und annotiert in verschiedener Weise solche digitalen Surrogate und besitzt damit einen komplexeren Inhalt sowie Struktur, die ebenfalls beschrieben werden müssen.*

Wie Abbildung 5.5 zeigt, werden digitale Surrogate in historische Korpora integriert. Der Zusammenhang zwischen den digitalen Surrogaten und der Quelle wird umfassend durch die TEI-Guidelines beschreibbar. Eine Quelle selbst ist durch die FRBR beschreibbar. Da ein Korpus mehr umfasst, als das, was die TEI und die FRBR beschreiben, und eine noch komplexere Struktur aufweisen, bedarf es eines weiteren Dateninhaltskonzepts für Metadaten für historische Korpora, der die unterschiedlichen Beschreibungskomponenten für Korpora als Ganzes sowie für die enthaltenen Texte und Annotationen berücksichtigt. Dafür liefert die hier vorliegende Arbeit einen Lösungsvorschlag in Form eines Metamodells für Korpusmetadaten (MKM) (Kapitel 6). Damit wird der Inhaltskonzept in Form eines Metamodells mit Hilfe der UML erstellt.

6 Metamodell für Korpusmetadaten

Das Metamodell für KorpusMetadaten (MKM)[126] modelliert Korpusmetadaten für historische Korpora, auf deren Basis die Akteurinnen und Akteure verschiedene Handlungen als Voraussetzung für eine Wiederverwendung ausführen können, wie in Odebrecht (2014) bereits motiviert wurde. Beispielsweise muss ein Korpus erst mit Hilfe von Metadaten und einer Metadatensuche gefunden werden (Handlung, Kapitel 4), bevor man es wiederverwenden kann (Szenarien, Kapitel 3). Für eine Wiederverwendung sind viele Aspekte der Korpuserstellung relevant, wie Kytö (2011) zeigt:

> It is important that compilers of future historical corpora pay attention to the above problems and that they document their compilation decisions in clear terms in user guides, corpus manuals and like material that will accompany the release versions of the corpora. [...] For later verification purposes, it is necessary for the respective corpus file or manual to contain bibliographical reference information on the specific copy used for the corpus. Overall, assessing the reliability and validity of source texts as evidence of language use from the past periods is of prime importance to any historical corpus compilation project. (Kytö 2011: 230-231)

Es wurde bereits zeigt, dass die korpuslinguistische Definition von Korpus mitsamt ihren Architektureigenschaften auf andere nicht-linguistische Korpora übertragbar ist. Damit stellen sich alle Korpora auch der schwierigen Definition von `Text` (Kapitel 2) und der Unterscheidung zwischen Quellen und ihren digitalen Surrogaten (Abschnitt 2.7.1, Kapitel 5). Weiterhin müssen die Eigenschaften der Korpusarchitekturen berücksichtigt werden, wie die verschiedenen Annotationskategorien und -konzepte sowie Tokenisierungen und Formate.

Die besondere Herausforderung besteht darin, ein Metamodell mit einem überfachlichen Geltungsanspruch zu entwickeln, um eine interdisziplinäre Wiederverwendung von Korpora zu ermöglichen. Die Auffassung, dass eine überfachliche Nutzung von

[126] Englisch ‚Metamodel for Corpus Metadata – MCM'.

Forschungsdaten auf Grund der theoretischen Perspektiven der unterschiedlichen Fächer kaum möglich sein soll (Sahle und Kronenwett 2013: 82), greift zu kurz. In Kapitel 1 wird eine Wiederverwendung von Korpora auf Grundlage gemeinsamer empirischer Grundlagen motiviert. Dies wird auch durch flexible Korpusarchitekturen unterstützt, die viele, konfligierende Interpretationen in einem Korpus ermöglichen (Kapitel 2). Kapitel 3 gibt bereits einige Beispiele, wie Korpora wiederverwendet wurden und zeigt allgemeine Szenarien der Wiederverwendung auf. Wie in Kapitel 5 gezeigt wurde, existieren bislang einige Ansätze, deren ganz konkretes Ziel eine solche überfachliche und einheitliche Beschreibung für die Wiederverwendung von Ressourcen ist.

Durch die Verortung von Korpuseigenschaften in Bezug auf den Forschungsprozess (theoretische Perspektive) und durch die Berücksichtigung des Forschungsdatenprozesses (datenzentrische Perspektive) lassen sich theoretisch-fachspezifische und einheitlich-überfachliche Konzepte identifizieren. Letztere werden hier als technisch-abstrakte Eigenschaften beschrieben, die u.a. die Korpusarchitektur als Ganzes betreffen, erstere sind z. B. Annotationskategorien und Textdefinitionen. Diese verschiedenen Eigenschaften sind mit den unterschiedichen Beschreibungskomponenten in Abbildung 5.1 zusammengefasst. Eine auf diesen Aspekten beruhende Korpusdokumentation, wie sie hier mit dem MKM vorgeschlagen wird, ist eine der Voraussetzungen für eine Wiederverwendung von Korpora.

Eine technisch-abstrakte Modellierung kann die verschiedenen Korpusarchitekturen (nicht nur für linguistische Korpora) abbilden und, wie Lüdeling (2012) erklärt:

> On an abstract technical level, there are no categorical differences between a large corpus for a well-researched language with many resources and a standardized orthography and a corpus of an endangered language or small variety without codified standards: In both cases one needs to represent a source text and annotations to it. (Lüdeling 2012: 32)

Dieser Ansatz wird auch in der vorliegenden Arbeit verfolgt: Die Dokumentation von Korpora stützt sich auf eine technisch-abstrakte Beschreibungsebene, die theoretische und fachspezifische Aspekte von Forschungsdaten zwar berücksichtigt, aber nicht darauf aufbaut.

Die Heterogenität der Datenlage stellt für Metadaten und deren Funktionen eine Herausforderung dar. So stellen beispielsweise Jensen et al. (2011) in Bezug auf Messdaten folgende Anforderungen an Metadaten:

> Für die Dokumentation von Forschungsdaten sind – neben der Beschreibung der Genese der Daten durch das Projekt und dessen Erhebungsinstrumente – die Bedeutung der Daten selbst zentral. Für die Reproduktion einer Datenanalyse oder die Erstellung von Sekundäranalysen muss die Bedeutung von Messreihen und -werten klar definiert werden. (Jensen et al. 2011: 86)

Übertragen auf die Eigenschaften der historischen Korpora müssen deren Erzeugung, die Tools zu ihrer Erstellung, sowie die verwendeten Annotationen klar mit Hilfe von Metadaten dokumentiert werden, um eine Wiederverwendung zu ermöglichen.

Das MKM wird mit der Unified Modeling Language (UML) modelliert, daher werden in Abschnitt 6.1 die verwendeten Modellierungsmechanismen erklärt. In Abschnitt 6.2 wird die Drei-Ebenen-Modellierung genauer beschrieben, in die das MKM eingebettet ist. Darauf folgt dann in Abschnitt 6.3 die genaue Beschreibung des Metamodells.

6.1 Modellierung nach UML

Modellierung wird in allen geisteswissenschaftlichen Fächern auf verschiedene Weise genutzt (Heßbrüggen-Walter 2016: 164).[127] Modellierung ermöglicht ganz allgemein eine vereinfachte, idealisierte Repräsentation von Wissen:

> [...] a model is by nature a simplified and therefore fictional or idealized representation [...]. (Cartwright 1983: 158 zitiert nach McCarty 2004: 255)

Zipser (2009) motiviert im Rahmen der Entwicklung eines Metamodells für linguistische Korpusdaten den Einsatz von Modellierung so:

> Durch ein Modell wird versucht, einen bestimmten Ausschnitt der Realität vereinfacht darzustellen. Welche Ausschnitte der Realität in einem Modell enthalten sind, hängt von dem Zweck eines Modells ab. Diese Abstraktion ist ein wichtiger Schritt, um die Realität besser zu verstehen und greifbarer zu machen. Ein Modell im Sinne der modellbasierten Entwicklung besitzt immer ein Metamodell, durch das es definiert wird. (Zipser 2009: 12-13)

[127] Auch Abstraktionen und Kategorisierungen wie in Kapitel 2 dargestellt sind Modellierungen.

Ein Modell für Korpusmetadaten versucht dann, Zipser (2009) folgend, einen bestimmten Ausschnitt von Eigenschaften eines historischen Korpus, die für den Zweck der Wiederverwendung relevant sind, vereinfacht darzustellen. Ein Metamodell für Korpusmetadaten definiert dann dieses Modell. Dazu wird die Unified Modeling Language (UML) als eine etablierte Modellierungssprache genutzt (ISO und IEC 2012).[128]

Diese Modellierungssprache ermöglicht eine sehr abstrakte Perspektive, Metadaten unabhängig von Fachgebieten und Formaten zu entwerfen und zu analysieren:

> Die Unified Modeling Language (UML) dient zur Modellierung, Dokumentation, Spezifikation und Visualisierung komplexer Softwaresysteme, unabhängig von deren Fach- und Realisierungsgebiet. (Rupp et al. 2005: 12)

Mit der UML ist es möglich, Objekte wie Gegenstände, Personen und Begriffe zu modellieren und dabei deren Eigenschaften und Beziehungen untereinander formal herauszuarbeiten. In dem hier verwendeten Klassenmodell werden Objekte als Klassen mit Attributen und Assoziationen untereinander abgebildet. Ein Klassenmodell fasst Objekte – also Personen, Dinge oder Konzepte – als Klassen abstrahierend zusammen, weist diesen **Klassen** bestimmte Eigenschaften in Form von *Attributen* zu und bildet Beziehungen (`Relationen`) zwischen Klassen ab (Rupp et al. 2005: 96). Klassen, Attribute, Objekte und Werte werden im Text jeweils anders hervorgeboben: **Klasse**, *Attribut*, Objekt (Paraphrasierung *ein Objekt der Klasse X*), *Wert*.

[128] Vgl. Wiese (2015) und Rupp et al. (2005) für eine Einführung in UML.

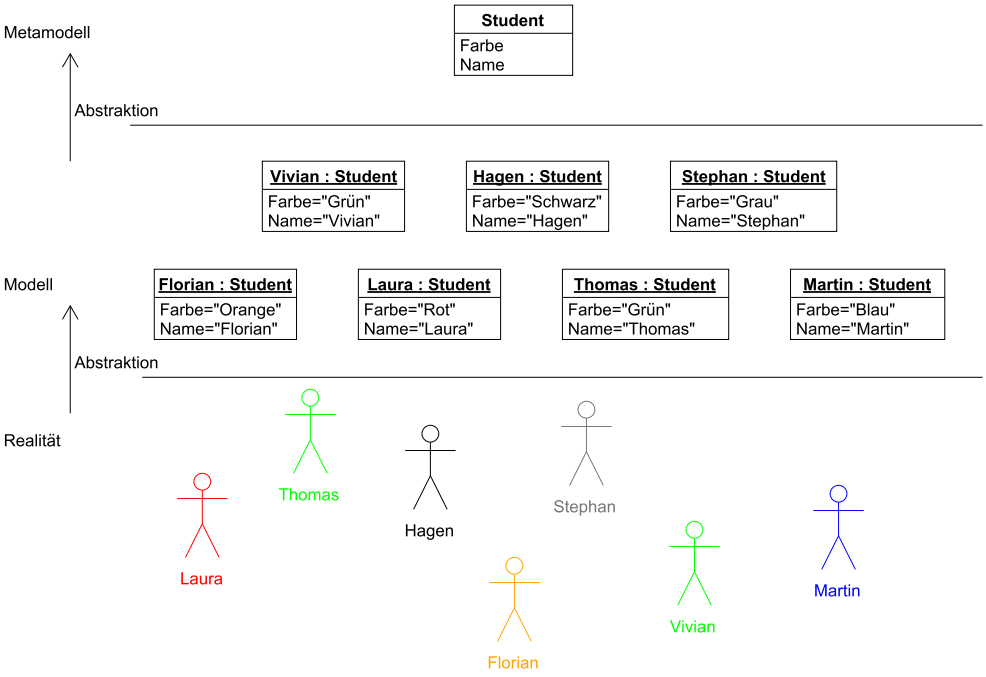

Abbildung 6.1: *Beispiel einer Modellierung. In der Realität gibt es eine Menge an Studentinnen und Studenten. Für eine bestimmte Forschungsfrage werden diese ganz abstrakt dargestellt bzw. modelliert. Dabei werden nur zwei der vielen Eigenschaften von Studierenden erfasst: Farbe und Name. Die Metamodellebene fasst gleichartige so modellierte Objekte und deren gemeinsame Eigenschaften zusammen: die Klasse* ***Student*** *mit den Attributen Farbe und Name.*

In Abbildung 6.1 werden drei Ebenen dargestellt. In der untersten Ebene *Realität* ist eine Menge von Studentinnen und Studenten mit Hilfe der Strichfiguren dargestellt. Die mittlere Ebene *Modell* enthält Objekte, die die realen StudentInnen mit zwei ihrer Eigenschaften repräsentieren. Eine Studentin oder ein Student wird jeweils mit Farbe und Namen beschrieben: Es gibt in der Modellebene ein Student mit der Farbe *Orange* und dem Namen *Florian*, ein Student mit der Farbe *Rot* und dem Namen *Laura* etc.[129] In der Realität können StudentInnen darüber hinaus noch wesentlich mehr Eigenschaften wie z. B. Studienfach, Fachsemester oder Studiengang besitzen, die hier nicht mit modelliert werden. In einem Modell wird daher aus der Realität ein Ausschnitt für einen bestimmten Zweck abstrahiert. In der obersten Ebene findet ein weiterer Abstraktionsschritt statt: Hier werden alle gleichartigen

[129] Diese Eigenschaften sind einzeln und in ihrer Kombination willkürlich gewählt und dienen nur Illustrationszwecken.

Objekte und die gemeinsamen Eigenschaften in Klassen und Eigenschaften – *Attributen* – abgebildet.

Die UML ist eine Notation für solche Modellierung. Sie gibt hingegen nicht vor, *was* auf *welche Weise* modelliert wird. Klassen mit Attributen und ihren Relationen untereinander werden einheitlich in dieser Arbeit nach UML dargestellt. Eine Klasse wird als ein Rechteck dargestellt, das in zwei weitere Rechtecke horizontal unterteilt ist, wobei die erste Zeile den Namen der Klasse enthält. In der zweiten Zeile befinden sich die Attribute und Werttypen der Klassen. Attribute besitzen einen Namen und einen Wert, der mit einem bestimmten Datentyp gekennzeichnet werden kann: z. B. *String* (Zeichenkette) oder *Integer* (numerischer Wert). [130]

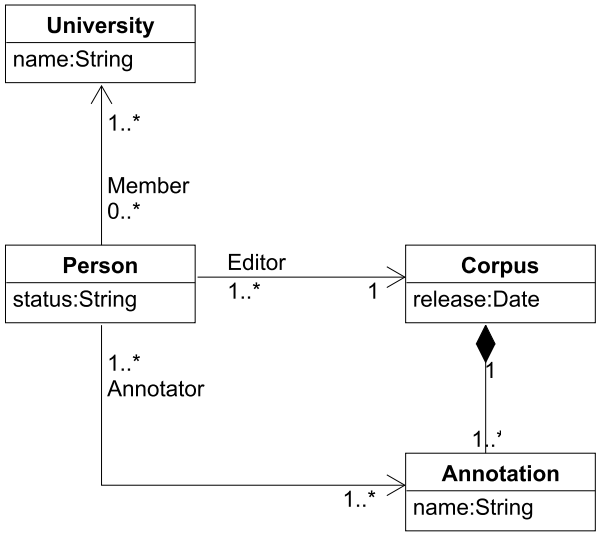

Abbildung 6.2: *Beispiel für Notationen der UML. Klassen wie* ***Corpus****,* ***Person****,* ***Annotation*** *und* ***University*** *werden in Rechtecken dargestellt und fett gedruckt. Ihre jeweiligen Attribute stehen im unteren Teil des Rechtecks und erhalten Werttypen. Die Relationen zwischen den Klassen wird durch verschiedenen Verbindungen zwischen den Rechtecken ausgedrückt. Die Relationen können durch die Angabe von Bezeichnungen, Multiplizität und Rollen qualifiziert werden.*

Die Klasse **Corpus** in der Abbildung 6.2 besitzt ein Attribut *release* mit einem Werttyp *Date*.[131] Das heißt, jedes Objekt der Klasse **Corpus** besitzt die Eigenschaft

[130] Attributtypen können selbst wiederum Klassen sein. Attributwerte sind dann Objekte (Instanzen von Klassen). In dieser Arbeit werden nicht alle Merkmale, Assoziationen und Funktionen der UML genutzt.

[131] Die Modellierung in UML erfolgt in dieser Arbeit grundsätzlich auf Englisch.

release; das Veröffentlichungsdatum eines Korpus wird mit *Date* angegeben.[132] Die Klasse **Person** besitzt ein Attribut *status* mit einem Typ *String*. Das heißt, jedes Objekt der Klasse **Person** besitzt die Eigenschaft *status*. Im Wertebereich von *status* könnte sich beispielsweise „Professor“ oder „Mitarbeiter“ befinden. Die Klasse **Annotation** besitzt ein Attribut *name* mit einem Typ *String*. Das heißt, jedes Objekt der Klasse **Annotation** besitzt die Eigenschaft *name*; der Name der Annotation wird angegeben. Die Klasse **University** besitzt ein Attribut *name* mit dem Werttyp *String*, der Name des Objektes der Klasse **University** wird angegeben.

Zwischen den einzelnen Klassen können Relationen bestehen, die in verschiedenen Arten von `Assoziationen` modelliert werden. In Abbildung 6.2 zeigt ein Pfeil von der Klasse **Person** zur Klasse **Annotation**. Damit werden Person zu Annotation zugeordnet. Die an dem jeweiligen Ende angegebenen Intervalle zeigen an, wie viele Male ein bestimmtes Element in der Realität instanziiert werden muss oder darf (Rupp et al. 2005: 75). Demnach können ein oder mehrere Person (1..*) einem oder mehreren Annotation (*1..**) zugeordnet werden (`Multiplizität`). Mit der Angabe eines Attributes *Annotator* an dem Ende der Relation von der Klasse **Person** kann die Assoziation weiter qualifiziert werden, so dass Person mit der Eigenschaften, *Annotator* zu sein, Annotation zugeordnet wird. Diese Angabe eines solchen Attributs wird in dieser Arbeit *mit der Rolle Annotator* paraphrasiert: Jedes Objekt der Klasse **Person** wird in der Rolle *Annotator* mit einem oder mehreren Objekten der Klasse **Annotation** assoziiert.

Die Klasse **Person** ist ebenfalls mit der Klasse **Corpus** assoziiert. In diesem Fall ist die Assoziation mit einer anderen Rolle (*Editor*) versehen. So können *1 ..** (ein bis unendlich viele) Korpora mit ebenfalls *1..** (ein bis unendlich vielen) Personen in der Rolle *Editor* assoziiert werden.

Zwischen der Klasse **Corpus** und der Klasse **Annotation** wird die Beziehung `Komposition` mit einer linearen, gerichteten Linie mit einer ausgefüllten Raute ausgedrückt (Rupp et al. 2005): „Sie drückt [...] die physische Inklusion der Teile im Ganzen aus. Teile und Ganzes bilden eine Einheit, deren Auflösung durchaus die Zerstörung des Ganzen zur Folge haben kann.“ (Rupp et al. 2005: 145) Wird das Aggregat gelöscht, werden auch seine Bestandteile gelöscht. Das Aggregat definiert den Rahmen und damit den Lebenszyklus der Komponenten. Die Komposition weist dann mit der Angabe von Multiplizität `variable` Bestandteile auf, hier *1..**. Die Bestandteile, Objekte der Klasse **Annotation**, können hinzugefügt oder entfernt werden, bevor das Aggregat, ein Objekt der Klasse **Corpus**, gelöscht wird. Die Auswahl

[132] Der Werttyp *Date* kann beispielsweise nach einer Norm definiert werden.

der Multiplizität ist bei der Komposition beschränkt; ein Ganzes darf über mehrere Teile verfügen, ein Teil aber nicht über mehrere Ganze (Rupp et al. 2005: 146).

Die Komposition drückt Folgendes in diesem Fall aus: Objekte der Klasse **Annotation** sind physische Bestandteile eines Objektes der Klasse **Corpus**. Ein *1..** Objekt der Klasse **Annotation** kann nicht ohne *ein* Objekt der Klasse **Corpus** existieren. So wird beschrieben, dass ein Korpus aus mehreren Annotationen besteht.

Damit sind die grundlegenden Modellierungsmechanismen, die in dieser Arbeit verwendet werden, vorgestellt. Im Weiteren wird das MKM in Abschnitt 6.2 in seiner Einbettung in die Drei-Ebenen-Modellierung und in Abschnitt 6.3 im Detail vorgestellt.

6.2 Drei-Ebenen-Modellierung für Korpusmetadaten

Die Modellierung der Metadaten von historischen Korpora sind vergleichbar mit der Modellierung der StudentInnen in Abbildung 6.1. Das MKM wurde entlang existierender Korpora wie RIDGES, KAJUK, GerManC oder dem Referenzkorpus Altdeutsch entwickelt.[133] Über deren gemeinsame Eigenschaften, die in Kapitel 2 detailliert vorgestellt wurden, wird dann abstrahiert, so dass die Modellierung möglichst allgemein für historische Korpora gehalten wird, um so auch auf weitere unbekannte Korpora anwendbar zu sein. So wird auch hier über einer Menge an gleichwertigen Objekten abstrahiert und deren Eigenschaften und Relationen untereinander in einem Metamodell beschrieben (vgl. Abbildung 6.3). Das MKM ist ein phänomenologisches Modell im Sinne von Heßbrüggen-Walter (2016), dass „lediglich empirische Daten und ad-hoc-Hypothesen [umfasst], wenn eine Theorie des Gegenstandsbereichs noch nicht zur Verfügung steht“ (Heßbrüggen-Walter 2016).

[133] Für eine Liste der Korpora, die mit dem MKM abgebildet werden können, vgl. `http://www.laudatio-repository.org/repository/view` (besucht am 13.01.2017).

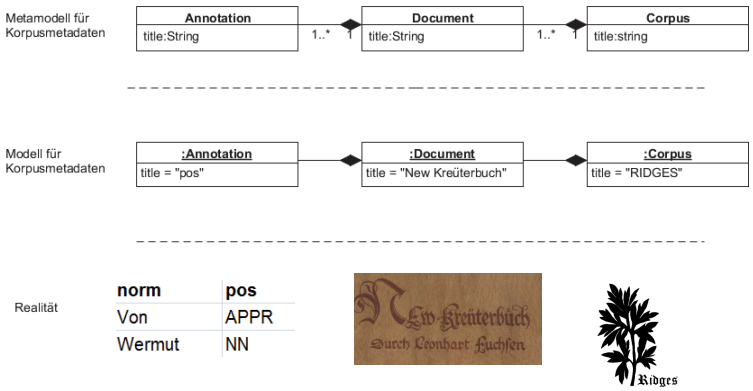

Abbildung 6.3: *Drei-Ebenen-Modellierung. Reale ganzheitliche Dinge wie eine Wortartenannotation pos, das Buch New Kreüterbuch oder das RIDGES-Korpus sind in der unteren Ebenen dargestellt, hier mit jeweils einem Bild. In der Realität enthält das RIDGES-Korpus ein solches Dokument und dieses wiederum enthält pos-Annotationen. Diese Dinge werden in der nächsten Ebene als Objekt repräsentiert und ausschnitthaft mit der Eigenschaft, einen Titel zu besitzen, modelliert. Die obere Ebene des Metamodells abstrahiert über alle gleichartigen Objekte und deren gemeinsamen Eigenschaften.*

Die Drei-Ebenen-Modellierung in Abbildung 6.3[134] repräsentiert in unterschiedlichen Abstraktionsgraden objektorientierte Korpusmetadaten. Das Korpus RIDGES in der realen Welt enthält z. B. Dokumente wie das *New Kreüterbuch* und eine bestimmte Wortartenannotation (vgl. Kapitel 2). Deren vielfältige Eigenschaften können jeweils mit Metadaten beschrieben werden. In der nächsten Ebene werden diese Objekte dann ausschnitthaft mit der Eigenschaft, einen Titel zu besitzen, modelliert. Zusätzlich zeigt die Assoziation (Linie mit ausgefüllter Raute) eine Komposition zwischen einem Objekt der Klasse **Corpus** (z. B. RIDGES) und einem Objekten der Klasse **Document** an. Als nächsten Abstraktionsschritt werden alle gleichwertigen Objekte zu Klassen zusammengefasst und deren gemeinsame Eigenschaften und Relationen untereinander in einem Metamodell beschrieben. So wird jedes Objekt der Klasse **Annotation** mit einen Titel beschrieben, genauso wie jedes Objekt der Klassen **Document** und **Corpus**. Weiterhin wird in dieser Ebene auch abstrahiert, dass ein Objekt der Klasse **Corpus** aus einem oder mehreren Objekten der Klasse

[134] Bild für die Annotation aus `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/v5/newkreuterbuch_1543_fuchs.xlsx`, Bild für das Buch aus `urn:nbn:de:bvb:12-bsb00017437-8`, Bild für RIDGES aus `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt`, (alle besucht am 08.01.2017).

Document besteht. Ein Objekt der Klasse **Document** besteht wiederum aus einem oder mehreren Objekten der Klasse **Annotation** (Abschnitt 6.3).

Das MKM ist ein Metamodell für Metadaten von Korpora, kein Metamodell für Korpusdaten wie beispielsweise das Metamodell *Salt* (Zipser und Romary 2010). Letzteres befasst sich mit der Abbildung von Annotationen wie sie z. B. in TEI-XML (TEI Consortium 2015) als Format für kritisch digitale Editionen, in EXMARaLDA (Schmidt und Wörner 2009) als Format für Dialogsprachdaten oder in TIGER-XML (Mengel und Lezius 2000) als Format für syntaktisch annotierte Korpora instanziiert sind:

> Salt unifies the concepts of these formats e.g. common timeline, multiple layers of annotation etc. and represents them in a common model. Salt is a model for representing the underlying organization of linguistic data, and as such, does not take into consideration their underlying semantics. Furthermore, Salt is independent of specific linguistic theories or analyse [sic!]. (Zipser und Romary 2010: 4)

Das Metamodell für Annotationsdaten befasst sich mit der einheitlichen Abbildung von verschiedenen Konzepten wie Zeitlinien oder Mehrebenenannotation, die in unterschiedlichen Formaten existieren. Hier geht es darum, die eigentlichen Annotationen eines Korpus abzubilden. Das MKM befasst sich hingegen mit der Beschreibung von Annotationsdaten und berücksichtigt auch korpusexterne Elemente. So werden Metadaten für Korpora, deren Dokumente und Annotationen in Verbindung mit Personen, Institutionen und den historischen Vorlagen modelliert.

6.3 MKM

Das MKM ist eine Modellierung, Dokumentation, Spezifikation und Visualisierung komplexer Metadaten von dokumentbasierten (historischen) Korpora unabhängig vom Format oder Fachgebiet, in dem sie entstanden sind.

Das Metamodell begreift vergleichbar mit Abney und Bird (2011: 125) das Korpus als eine Summe der Arbeitsschritte, die zu dessen Erstellung angewandt wurden. Aus den auch in Abschnitt 2.7 beschriebenen Korpora wird ein umfassendes Bild über diejenigen Eigenschaften entwickelt, welche sich Korpora häufig teilen. Folgende Annahme gilt: Alle Korpora teilen sich gemeinsame, technisch-abstrakte Eigenschaften unabhängig von ihrer durch Annotationen und Kategorien implementierten Forschungsfrage oder ihrem fachspezifischen Blick auf Konzepte wie *Primär-* oder

Sekundärtext. Es sollen abstrakte Korpuseigenschaften in Bezug auf beispielsweise die Tokenisierung, die Transkription und Annotationen einheitlich im Metamodell abgebildet werden. Spezifische Formen einer Tokenisierung, wie satz- oder wortweise Tokenisierung, können somit durch das gleiche Metamodell beschrieben werden.

Die folgende Frage ist zentral für die Entwicklung des MKM:

- Welche Eigenschaften von historischen Korpora sind relevant für deren Erschließung durch Dritte zum Zweck der Wiederverwendung?

Konkreter stellen sich die folgenden Fragen in Bezug auf die zu beschreibenden historischen Korpora:

- Welche Eigenschaften haben Korpora aus unterschiedlichen Fächern gemeinsam?
- Wie können oder sollen fachspezifische Interpretationen berücksichtigt werden?

Diese Fragen können unterschiedlich beantwortet werden, je nachdem, welches Wiederverwendungsszenario aktuell unterstützt werden soll (vgl. Kapitel 3). Um ein Korpus wiederverwenden zu können, müssen vorher auf Metadaten basierte Handlungen durchgeführt werden (vgl. zum Zusammenhang von Wiederverwendungsszenario und Handlung Abbildung 4.4): Nehmen wir an, ein Korpus soll mit weiteren historischen Texten erweitert werden (Szenario 4 Größenanreicherung). Durch Erschließung eines Korpus mit Hilfe von technischen Metadaten (Abschnitt 4.3) aus einer produktorientierten Perspektive (Abschnitt 4.4) stellen sich folgende Fragen: Wurde eine bestimmte Annotationsebene automatisch oder manuell erstellt? Welche Tags wurden in einer Annotationsebene vergeben? In welchem Format wurde annotiert? Wurden die Annotationen überprüft? Die Antworten sind wichtig, um die gleichen Annotationsebenen auf neue Texte für die Erweiterung des Korpus anwenden zu können. Nicht relevant dafür sind beispielsweise Informationen über gelöschte beziehungsweise nicht vorhandene Tags, Zeitpunkte der Annotation oder Korrektur. Wenn man beispielsweise herausfinden möchte, welche Wortartentags in einem Korpus verwendet werden, ist eventuell die Information, wer diese vergeben hat, weniger relevant. Für eine Information über die Herkunft dieser Annotation sind jedoch die Angaben, wer (oder welches Tool) annotiert hat, wiederum relevant. Korpora besitzen damit viele Eigenschaften, die nicht alle (im gleichem Maße) relevant für jeden Zweck sind.

In Bezug auf die Funktion und die Handlungsszenarien der Metadaten stellen sich folgende Fragen:

- Welche Informationen über Korpora helfen Akteurinnen und Akteuren, diese fach-, format und erstellerunabhängig zu erschließen?
- Wie müssen diese Informationen idealerweise strukturiert sein?
- Welche Bezugspunkte sind dafür hilfreich?

Das MKM legt drei Konzepte fest, die zentral für die Beantwortung der obigen Fragen und als Bezugspunkte für die verschiedenen Eigenschaften der Korpora verstanden werden. In Abbildung 6.3 sind konkrete Beispiele dafür gegeben: Eine Annotation kann als Attribut-Wert-Paar verstanden werden (z. B. *pos*). Ein Dokument (z. B. *New Kreüterbuch*) besteht aus mindestens einer Annotation. Eine Korpus (z. B. *RIDGES*) besteht aus mindestens einem Dokument. Diese drei Konzepte werden in den drei Klassen **Corpus**, **Document** und **Annotation**, die in einem Kompositionsverhältnis zueinander stehen, modelliert. Abbildung 6.4 zeigt einen vereinfachten Ausschnitt des MKM, in dem keine weiteren Attribute der drei Klassen oder anderer Klassen und deren Relationen untereinander abgebildet sind.

Abbildung 6.4: *Vereinfachte Darstellung der zentralen Klassen im MKM. Ein Objekt der Klasse **Corpus** enthält ein oder mehrere Objekt der Klasse **Document**. Ein Objekt der Klasse **Document** enthält ein oder mehrere Objekte der Klasse **Annotation**.*

Die Objekte der Klasse **Document** existieren nicht unabhängig von den Objekten der Klasse **Annotation**. Im Umkehrschluss existieren für das MKM Objekte der Klasse **Document** nur, wenn es mindestens ein Objekt der Klasse **Annotation** gibt. Dasselbe gilt für die Assoziation von der Klasse **Corpus** und der Klasse **Document**. Objekte der Klasse **Corpus** existieren nicht unabhängig von Objekten der Klasse **Document**.

Mit einer solchen grundlegenden Struktur und der angebenden Multiplizität gilt dann für Objekte der jeweiligen Klassen: ein Document kann mehrere Annotation beinhalten. Damit sind die Annotation variable Komponenten eines Document (das Ganze), die hinzugefügt ausgetauscht oder (bis auf ein (1) Annotation) gelöscht werden können (vgl. auch Abschnitt 2.7.3). Wenn ein Document gelöscht wird, werden

auch alle Annotation gelöscht. Dieses gelöschte Document und alle Annotation können nicht mehr mit Metadaten beschrieben werden. Wenn ein Document mehrere Annotation enthält und ein Annotation davon gelöscht wird, dann ist das Document immer noch ein Aggregat aus ein oder mehreren Annotation, dadurch dass *1..** Annotation Bestandteile eines Document sind. Ein Dokument besteht also mindestens aus einer Annotation. Eine solche Annotation kann beispielsweise eine Transkription oder Normalisierung einer historischen Vorlage sein.

Gleiches gilt für die Komposition zwischen der Klasse **Corpus** und der Klasse **Document**, die ebenfalls mit einer Multiplizität angegeben ist: Wenn ein Corpus mehrere Document enthält, dann existiert es immer noch, wenn ein oder mehrere Document hinzugefügt, geändert oder (bis auf ein (1) Document) entfernt werden.

Abbildung 6.5 zeigt die Relationen der Klassen anhand eines Ausschnitt von RIDGES. Hierbei gilt, ein Objekt ist eine Instanz einer Klasse.

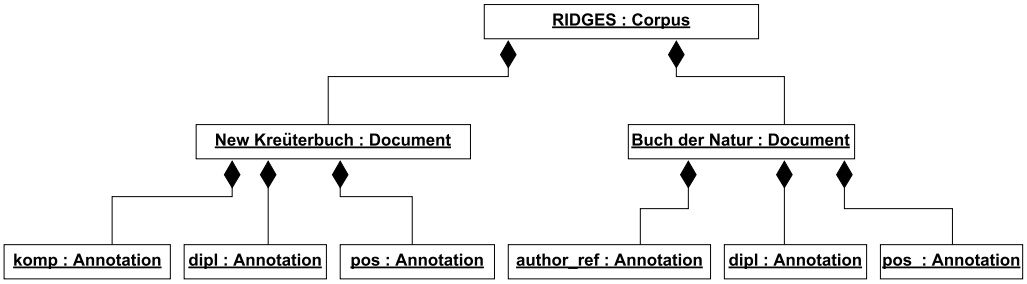

Abbildung 6.5: *Instanzenmodell RIDGES (Ausschnitt). Das RIDGES-Korpus ist eine Instanz der Klasse* ***Corpus****. Buch der Natur und New Kreüterbuch sind Instanzen der Klasse* ***Document*** *und dipl, pos, author_ref und komp Instanzen der Klasse* ***Annotation****. Das RIDGES-Korpus enthält zwei Dokumente: Buch der Natur und New Kreüterbuch. Das Dokument Buch der Natur enthält die Annotationen dipl, pos und author_ref. Das Dokument New Kreüterbuch enthält ebenfalls dipl, pos sowie komp.*

Ein Korpus wie RIDGES (Version 5.0) besteht aus mehreren historischen Texten, die jeweils als Dokument beschrieben sind, wie beispielsweise Auszüge aus dem *New Kreüterbuch*[135] und dem *Buch der Natur von Megenberg*[136]. Diese wiederum enthalten u. a. Annotationen in Form von Transkriptionen und Wortarten oder Autoren. So existieren die Dokumente *Megenberg* und *New Kreüterbuch* aus dem Korpus RIDGES aus der Sicht des MKM nicht jeweils unabhängig von ihren Annotationsebenen: Das

[135] New Kreüterbuch / Leonhart Fuchs (Basel, 1543) [S. 2-e4 (25 Seiten), 5221 dipl-Einheiten]
[136] Das Buch der Natur, Konrad von Megenberg (Augsburg, 1482), [S. NA (15 Seiten), 5215 dipl-Einheiten].

Dokument *Buch der Natur* besteht aus *dipl*, *pos*, *author_ref*. Das Dokument *New Kreüterbuch* besteht aus *dipl*, *pos* und *komp*.

6.3.1 Die Klasse Annotation

Diese Klasse bildet alle Metadaten für alle Annotationen in einem Dokument ab und beschreibt dabei Annotationskategorien und -konzepte inklusive Primärtext- oder Sekundärtextinterpretationen wie Transkriptionen oder Normalisierungen einheitlich. Durch die Heterogenität der Forschungsdaten und deren vielfältigen Annotationen kann die Frage nach einer Deutungshoheit bei der Ausweisung von Konzepten wie `Primärtext` oder der Abbildung von verschiedenen Bedeutungen von einzelnen Annotationsebenen auf eine gemeinsame Bedeutung nicht beantwortet werden und wird daher auch nicht im Modell verankert (Odebrecht 2015). Es werden alle unterschiedlichen Transkriptionsebenen in den verschiedenen Korpora (vgl. Abbildung 2.4, Abbildung 2.5 und Abbildung 2.6) genauso wie alle beispielsweise linguistischen Annotationen für Wortarten *pos* (Abschnitt 2.3.1) und nicht-linguistische Annotationen wie *author* (vgl. Abschnitt 2.3.2) einheitlich als Objekte der Klasse **Annotation** modelliert. Damit motiviert sich ebenfalls, warum ein Document eine Komposition aus einem oder mehreren Annotation modelliert wird. Ein oder mehrere Annotation können beispielsweise als eine Transkription oder Normalisierung den historischen Text (das Dokument) repräsentieren. Die Metadaten des Dokumentes werden dann in der Klasse **Document** modelliert (Abschnitt 6.3.2).

Abbildung 6.6 zeigt einen Ausschnitt des MKM mit der Klasse **Annotation** und ihren Attributen sowie weiteren Assoziationen zu anderen Klassen.

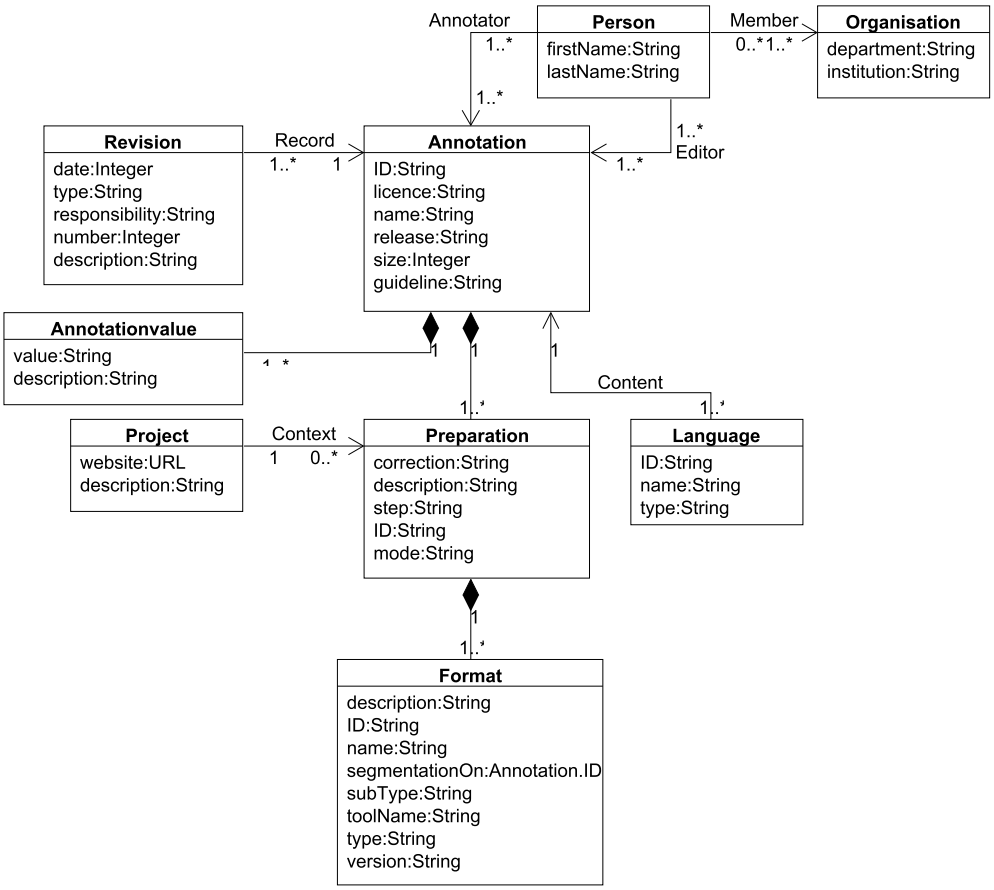

Abbildung 6.6: *Die Klasse* ***Annotation*** *und ihre Attribute sowie weitere Relationen zu anderen Klassen und ihren Attributen. Einem Objekt dieser Klasse werden Werte, Bearbeitungsschritte und die verwendeten Formate zugeordnet. Dazu werden jedem Objekt dieser Klasse noch Personen, die als Annotator oder Herausgeber der Annotation auftreten, sowie ein oder mehrere Sprachen zugeordnet.*

Die Informationen über die Autorenschaft (Annotatoren) und auch die Herausgeberschaft von Annotationen sind wesentliche administrative Metadaten, die Ansprechpartner und Verantwortliche identifizieren und eine Referenzierung mit Namensnennung ermöglichen. Annotationen werden einerseits Personen zugeordnet, die eine Rolle als Autor bzw. als Annotator besitzen, und andererseits Personen zugeordnet, die sie herausgeben (Klasse **Person** mit zwei Rollen *Annotator* und *Editor* sowie Klasse **Organisation**). Ein oder mehrere Person werden in ihrer Rolle als *Annotator* jedem Annotation zugeordnet, um so jeder Annotation einen Annotator

zuzuweisen. Damit können auch konfligierende Annotationen in einem Korpus ihren jeweiligen Annotatoren zugeordnet werden und für Dritte ist transparent, dass zwei verschiedene Forschungsvorhaben an einem Korpus durchgeführt worden sind. Ein Herausgeber einer Annotation muss nicht zwangsläufig auch Annotator sein. Damit wird jedem Objekt der Klasse **Annotation** ein oder mehrere Objekte der Klasse **Person** mit einer bestimmten Rolle zugeordnet. Nicht allen Objekten der Klasse **Person** müssen ein oder mehrere Objekte der Klasse **Organisation** zugeordnet werden. Beispielsweise ist nicht jede Person ein Mitglied einer Organisation oder Mitglied mehrerer Organisationen.

Ganz allgemein wird jedes Objekt der Klasse **Annotation** mit deskriptiven Metadaten durch ihre Attribute *name* und *guideline* beschrieben. Mit *guideline* wird eine Referenz auf eine entsprechende Annotationsrichtlinie abgebildet. Mit den Attributen *ID*, *size* und *release* derselben Klasse werden technische Metadaten zu Größe, Veröffentlichungsdatum und Version der Annotationen angegeben. Das Attribut *licence* gibt administrative Metadaten zur Lizenz von jedem Objekt der Klasse **Annotation**.

Um einen Überblick in die Bearbeitungsgeschichte einer Annotation zu ermögichen, werden jedem Annotation ein oder mehrere Revision zugeordnet, die die Revisionsgeschichte der vorherigen Versionen der Annotation beschreiben. Die Klasse **Revision** enthält die Attribute *date*, *type*, *number* und *description*, die technische Metadaten für ein Annotation zu vorherigen Versionen inkl. Datum, Nummer, Typ, Verantwortlichkeiten und einer Beschreibung abbildet.

Jede Annotation besteht unabhängig von den genutzten Kategorisierungssystemen oder -konzepten aus mindestens einem Wert, der wiederum eine Beschreibung erhalten kann. Dies wird im MKM so abgebildet, dass jedem Annotation mindestens ein Annotationvalue zugeordnet wird. Wenn ein Objekt der Klasse **Annotation** gelöscht wird, so werden auch seine Bestandteile (Annotationvalue) gelöscht. Die Annotationvalue sind durch die angegebene Multiplizitäten variable Bestandteile eines Annotation. Damit sind Annotationvalue variable Komponenten, die hinzugefügt, ausgetauscht, oder (bis auf ein (1) Annotationvalue) gelöscht werden können. Durch deren Attribute *value* und *description* werden deskriptive Metadaten zur Annotation angegeben. Damit werden alle Werte einer Annotation und deren Beschreibungen abgebildet. Diese Werte müssen jedoch nicht eins zu eins einer Annotationsrichtlinie entsprechen. Hier werden lediglich die tatsächlich im entsprechenden Korpus verwendeten Werte dargestellt. Nicht alle Werte einer Richtlinie müssen immer in einem Korpus annotiert sein. Auch Anpassungen bei der Verwendung von

Annotationswerten und deren Zuweisung sind möglich. Für die Erschließung (und Wiederverwendung) ist es auch hilfreich zu wissen, welche Dinge nicht annotiert wurden (weil sie nicht im Korpus vorkommen oder Teile der Annotationsrichtlinie nicht umgesetzt wurden). Dafür reicht z. B. eine Angabe, dass Wortarten annotiert worden sind, oder eine Nennung einer Richtlinie wie *Annotation nach dem STTS* als deskriptives Metadatum für die Beschreibung von Annotationen nicht aus.

Jede Annotation liegt in einer oder mehreren Sprachen vor. Beispielsweise können einige Annotationen und ihre Kategorien in Englisch und andere Annotationen und ihre Kategorien in Deutsch oder auch gemischt vorliegen, so dass jedem Objekt der Klasse **Annotation** noch ein oder mehrere Objekte der Klasse **Language** zugewiesen werden.

Jede Annotation wird durch eine oder mehrere Schritte erstellt (Forschungsdatenzyklus). Eine Annotation ist eine Komposition aus Bearbeitungsschritten. Dies wird im MKM so abgebildet, dass jede Annotation im Korpus (Annotation) nicht unabhängig von ihren Bearbeitungsschritten (Preparation) existiert. Die Angabe der Bearbeitungsschritte einer Annotation stellen technische Metadaten dar, die sich auf das fertige Korpus als Produkt aus Arbeitsschritten (produktorientiert) beziehen (Abschnitt 4.4).

Die Attribute der Klasse **Preparation**, nämlich *step*, *ID*, *mode*, *description* und *correction* geben technische und deskriptive Metadaten an, die jeden Schritt mit einer Referenz, Art der Bearbeitung und Korrekturverfahren sowie mit einer Beschreibung angeben. Das Attribut *step* gibt an, um welche Art der Bearbeitung einer Annotation es sich handelt, beispielsweise eine bestimmte Art Annotation oder Konvertierung. Jeder Schritt wird als administrative Information zum Kontext keinem, einem oder mehreren Projekten (Project) zugeordnet, in dem sie erstellt wurden.

Jede Annotation ist in irgendeiner Weise in einem oder mehren Formaten umgesetzt, die mit der Klasse **Format** modelliert werden. Technische Metadaten werden als Attribute *name*, *version* und *ID* modelliert und geben den Namen und die Version des Formats und eine Referenz an. Mit ihren Attributen *toolName* und *description* werden technische Metadaten zu dem genutzten Tool und dem konkreten Bearbeitungsvorgang angegeben. Damit können alle in einem Bearbeitungsschritt verwendeten Tools und deren Formate dokumentiert werden.

Annotationen können in mehreren Formaten vorliegen und durch unterschiedliche Bearbeitungsschritte erzeugt werden. Eine Bearbeitung der Annotation ist nicht unabhängig von dem oder den genutzten Formaten, wobei die Objekte der Klasse **Format** durch die angegebene Multiplizität variabel sind. Wird ein Bearbeitungsschritt

entfernt (oder rückgängig gemacht), der mit einem Objekt der Klasse **Format** assoziiert wird, dann entfällt auch die Dokumentation des Formats für die Annotation. Die Unterscheidung von Format und Preparation ist ebenfalls relevant, wenn entweder unterschiedliche Bearbeitungsschritte in demselben Format oder gleiche Bearbeitungsschritte in unterschiedlichen Formaten vollzogen werden. Wenn ein Korpus in mehreren Formaten unabhängig voneinander annotiert wird, wird dies ebenfalls durch das MKM abgebildet. Annotiert werden kann beispielsweise automatisch in einem Format, manuell in einem anderen. Ein Bearbeitungsschritt wie Konvertieren kann eine Annotation von einem Format (aus einem ersten Bearbeitungsschritt) in ein weiteres Format überführen. Somit ist dieses Format dann Teil des Konvertierungsschritts (vgl. Abbildung 6.7).

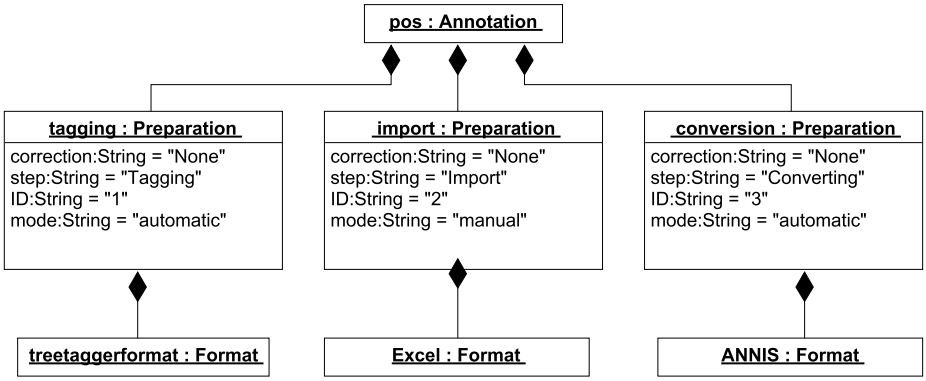

Abbildung 6.7: *Bearbeitungsschritte einer Annotation (verkürzte Darstellung). pos ist eine Instanz der Klasse **Annotation**. Diese Annotation wird in einem ersten Schritt automatisch erzeugt (tagging : Preparation). In einem weiteren Schritt wird diese Annotation mit import : preparation in Excel eingefügt. In einem nächsten Schritt wird diese Annotation in ein weiteres Format konvertiert (conversion : preparation).*

Abbildung 6.7 zeigt eine verkürzte Darstellung einer Instanziierung der Klasse **Annotation**. Die Instanz der Klasse **Annotation** *pos* wird über mehrere Schritte hinweg mit verschiedenen Formaten automatisch erzeugt, manuell bearbeitet und automatisch konvertiert. Diese Annotation wird in einem ersten Schritt automatisch erzeugt (tagging : Preparation). In einem weiteren Schritt wird diese Annotation manuell in Excel importiert (import : preparation). In einem nächsten Schritt wird diese Annotation in ein weiteres Format konvertiert (conversion : preparation).

So wird ermöglicht, verschiedene Bearbeitungsweisen (z. B. automatisch oder manuell) in verschiedenen Formaten mit verschiedenen Kontexten in einer Korpusdo-

kumentation zu berücksichtigen:

> At present the situation appears to be particularly complex because of the variety of contexts and forms that syntactic information may take. Firstly, syntactic information may either be the result of an automatic parsing of textual data or may be manually generated as a component of an annotated corpus. Secondly, the organisation and actual complexity of syntactic information highly depends on both the application context and theoretical background of the project within which such data has been created. (Romary et al. 2015: 2)

Was hier Romary et al. (2015) für syntaktische Annotationen beschreiben, lässt sich auch allgemein auf Annotationen übertragen (vgl. Abschnitt 2.3). Die Arbeit mit Annotationen ist komplex und muss die Art der Erstellung einer Annotation und das Annotationsmodell (Organisation und Komplexität von Informationen) mit einbeziehen.

Die Informationen über Organisation und Komplexität der Annotationen in einem Korpus werden ebenfalls innerhalb der Klasse **Format** modelliert. Die Attribute *segmentationOn*, *type* und *subType* beschreiben die Struktur und den Bezug der Annotation pro Format und Bearbeitungsschritt mit technischen und deskriptiven Metadaten. *type* und *subType* können dabei die Annotationskonzepte und -arten angeben. Im Wertebereich von *type* können dann die Annotationskonzepte wie *Spannenannotation* oder *Baumannotation* liegen. Im Wertbereich von *subType* sind beispielsweise freiere, inhaltliche Zuordnungen wie *Lexikalisch* oder *MarkUp* möglich. So kann eine Annotation als syntaktische (mit *subType*) Baumannotation (mit *type*) beschrieben werden, die über die Objekte der Klasse **Annotationvalue** mit entsprechenden Werten ausgestattet ist.

Weiterhin wird angegeben, welche Annotation in einem Format eigenständig tokenisiert ist oder sich auf eine bereits vorhandene Tokenisierung einer anderen Annotation bezieht (*segmentationOn*). Ein Document muss mindestens ein Annotation mit einer eigenständigen Segmentierung enthalten, unabhängig davon ob fest oder flexibel tokenisiert worden ist. Weiterhin kann auch mehrfache Tokenisierung vorliegen (multiple Segmentierungen). Eine nicht eigenständige Segmentierung liegt dann vor, wenn der Wert von (*segmentationOn*) eine ID einer anderen Annotation im Korpus enthält. Dies ist eine wesentliche Angabe für die Beschreibung der Korpusarchitektur. Damit gibt es Annotationen, die eine eigenständige Tokenisierung besitzen, und

Annotationen, die sich auf eine solche beziehen. Dies kann auch von Format zu Format unterschiedlich umgesetzt sein.

dipl	ſie	zu	allerley	handt⸗	griff	zuverſtehen		/
clean	sie	zu	allerley	handtgriff		zuverstehen		/
norm	sie	zu	allerlei	Handgriffen		zu	verstehen	/
lemma	sie	zu	allerlei	Handgriff		zu	verstehen	/
pos	PPER	APPR	PIAT	NN		PTKZU	VVINF	$(
lb	lb				lb			
p	p							
pb	pb							

Abbildung 6.8: *Multiple Segmentierungen in RIDGES. Die Annotationen dipl, clean und norm besitzen jeweils eine eigenständige Segmentierung. Die anderen Annotationsebenen beziehen sich jeweils auf eine dieser Annotationen.*

Das Beispiel[137] zeigt eine Korpusarchitekur mit multipler Segmentierung (vgl. Abschnitt 2.7.2). Jede der Annotationsebenen ist eine Instanz der Klasse **Annotation**. Sie liegen durch einen Bearbeitungsschritt (conversion : Preparation) in dem Format ANNIS vor. In diesem Format (ANNIS:Format) erhalten jeweils die Annotationen *dipl*, *clean* und *norm* zum Attribut *segmentationOn* den Wert ihrer jeweilige ID (Attribut ID der Klasse **Annotation**), da sie jeweils eine eigenständige Tokenisierung besitzen. Die Annotation *lb* beispielsweise erhält für das Attribut *segmentationOn* den Wert *dipl*, weil diese als Spannenannotation (Format.*type*) auf der Tokenisierung von *dipl* basiert. Die Annotation *pos* hat ebenfalls keine eigenständige Segmentierung und bezieht sich auf *norm*. Dies wird an den Einheiten *handt⸗* und *griff* in *dipl* deutlich, wobei sich eine Spanne auf der *lb*-Ebene auf *handt⸗* (und den vorausgegangenen Einheiten) und eine andere Spanne auf *griff* (und den nachfolgenden Einheiten) bezieht. Die Annotationsebene *pos* bezieht sich auf die Tokenisierung der *norm*, da *zuverſtehen* in *clean* und *dipl* noch eine Einheit darstellt, in *norm* jedoch zwei Einheiten *zu* und *verstehen* existieren, die jeweils einen eigenen Wert (*PTKZU* und *VVFIN*)[138] auf *pos* erhalten.

Alle Ebenen *dipl*, *clean*, *norm*, *pos* und *lb* werden in dieser Arbeit als Objekte der Klasse **Annotation** einheitlich repräsentiert. Sie bestehen aus ein oder mehreren

[137] RIDGES Version 5.0, AlchymistischePractic_1603_Libavius, Treffer `https://korpling.german.hu-berlin.de/annis3/?id=6816d01f-15d1-432a-bb91-5fcee9cd85b8` (besucht am 14.01.2017).

[138] PTKZU steht für *zu* vor Infinitiv. VVIFN steht für finites Vollverb.

Werten und werden durch ein oder mehrere Bearbeitungsschritte erzeugt und liegen in einem oder mehren Formaten vor. Die Werte der *pos*-Ebene in RIDGES entsprechenden den verwendeten Tags des STTS, die Werte der *dipl*-Ebene historischen Wortformen oder Wortmorphemen oder Zeichen allgemein. Letztere Wertemenge stellt keine klar begrenzte Menge an Tags dar. Sie ist vergleichbar mit Wertemengen wie Lemmatisierungen, die wiederum unstrittig als Annotationen verstanden werden. Die Bearbeitung kann in allen Fällen durch ihre einzelnen Schritte (z. B. Erstellung, Korrektur, Konvertierung) und ihren Modus (z. B. manuell oder automatisch) beschrieben werden. Alle Annotationen, *pos* wie *lb* wie *dipl* können dann in einem Format oder mehren Formaten vorliegen. Durch die Angabe der Annotationskonzepte und der Annotation(en) sowie der Segmentierung kann jede Korpusstruktur abstrakt beschrieben werden. Damit können auch Parallelkorpora wie Anselm mit dem MKM beschrieben werden.

Auf diese Weise werden alle Annotationen einheitlich im MKM als Objekte der Klasse **Annotation** modelliert. Inwieweit beispielsweise die Ebenen *dipl*, *clean* und *norm* in RIDGES jeweils ein digitales Surrogat oder ein Primärdatum darstellen, oder eine eigenständige Expression oder jeweils Realisierungen eines historischen Werks sind (vgl. Abbildung 2.3), wird nicht mit der Klassen **Annotation** im MKM modelliert.

In wie fern bei der Erschließung eines historischen Korpus potentielle Kandidaten identifiziert werden können, um ihnen eine bestimmte Repräsentationsfunktion zuzuweisen, ist hingegen mit der Angabe der Segmentierung pro Format, pro Bearbeitungsschritt und Annotation im MKM möglich. Typische Kandidaten von Annotationen, die auf eine bestimmte Weise einen Text repräsentieren, besitzen häufig eine eigenständige Segmentierung. Viele, wenn nicht alle anderen Annotationen beziehen sich dann auf diese Annotation. Eine so modellierte Beschreibung durch die Eigenschaften der Segmentierung überlässt es den Nutzerinnen und Nutzern, ihre jeweiligen Konzepte von `Text` anzuwenden.

Durch die Angabe aller Bearbeitungsschritte in einem oder mehreren Formaten, die zum vorliegenden Korpus (Produkt aus mehreren Bearbeitungsschritten) geführt haben, wird eine umfassende technische Dokumentation durch und für Akteurin/Akteur 1 (Inititalerstellung) und Akteurin/Akteur 2 (Inititalbearbeitung) sowie die Nachvollziehbarkeit der Korpuserstellung für Akteurin/Akteur 3 (Drittbearbeitung) ermöglicht. Die Metadaten liefern relevante Informationen, um Korpora mit weiteren Annotationen in den vorhanden Formaten anreichern oder reduzieren zu können (Szenario 3 und Szenario 5). Die vermittelten Informationen über die

Korpusarchitektur helfen dabei. So wird dokumentiert, welche Annotationen welche Segmentierungen besitzen. Die Löschung einer Annotation mit einer unabhängigen Segmentierung (z. B. in Szenario 2 (Korrektur)) kann unter Umständen dazu führen, dass die gesamte Korpusarchitektur fehlerhaft wird.

Mit diesem Ausschnitt des MKM können die Annotationen von historischen Korpora und mit einfacher Tokenisierung oder multipler Segmentierung durch Metadaten einheitlich und erstellerunabhängig dokumentiert werden – Handlung 1 (Deskription). Alle Annotationen werden einheitlich mit der Klasse **Annotationen** in dem Metamodell abgebildet, unabhängig davon, ob sie feste Kategorisierungen wie Wortartentagsets, offene Kategorisierungen wie Lemmatisierungen, oder kommentierende oder freie Tags wie bei Transkriptionen oder Erklärungen beinhalten. Damit ist das MKM auch vielfältig einsetzbar. Alle Annotationen können durch ihren Namen, die enthaltenen Werte und deren Erklärungen inhaltlich beschrieben werden. Für Szenario 1 (Analyse) sind die Angaben, welche Annotationsebenen in einem Korpus vorhanden sind, naturgemäß relevant, da in korpusbasierten Studien Annotationen die Grundlage der Analyse bilden. Jede dieser Annotationen kann in einem oder mehreren Formaten erstellt, korrigiert und konvertiert werden. Das Ergebnis aus diesen Bearbeitungsschritten kann mit diesen technischen Metadaten (modelliert in Form der Klassen **Preparation** und **Format**) dokumentiert werden. Wenn Korpora in mehreren Formaten vorliegen, ist es relevant zu wissen, welche Formate welche Annotationen eines Korpus abbilden. Dies gilt auch, wenn Korpora mit neuen Annotationen versehen (Szenario 3) und selbst erweitert (Szenario 4) oder Annotationen gelöscht (Szenario 6 und Szenario 5) werden sollen.

6.3.2 Die Klasse Document

Die zweite zentrale Klasse **Document** existiert nicht unabhängig von der Klasse **Annotation**. Die Klasse bildet alle deskriptiven und strukturellen Metadaten ab, die die im Korpus repräsentierten historischen Dokumente (Vorlagen) beschreiben.

Abbildung 6.9 zeigt die Klasse **Document** mit ihren Attributen sowie weiteren Assoziationen zu anderen Klassen. Verkürzt dargestellt sind alle Klassen, die bereits vorgestellt wurden. Jedes Objekt der Klasse **Document** wird mit deskriptiven und strukturellen Metadaten zum Namen, der Größe und Referenz beschrieben (*name*, *size*, *ID*). Die Informationen über die Autorenschaft und Herausgeberschaft von Dokumenten sind wesentliche, in diesem Fall bibliographische und keine administrativen Metadaten, die Verantwortliche identifizieren und eine Referenzierung mit

Namensnennung ermöglichen. Die Instanzen von der Klasse **Person**, die einem Dokument zugeordnet sind, sind daher andere als die, die einer Annotation zugeordnet sind. Der Autor eines Dokuments wie z. B. Leonhart Fuchs (*New Kreüterbuch*) kann beispielsweise nicht ein Annotator sein und der Annotator nicht der Autor des Dokumentes *New Kreüterbuch*. Damit geben in diesem Fall die Objekte der Klasse **Person** bibliographische Metadaten zum Dokument.

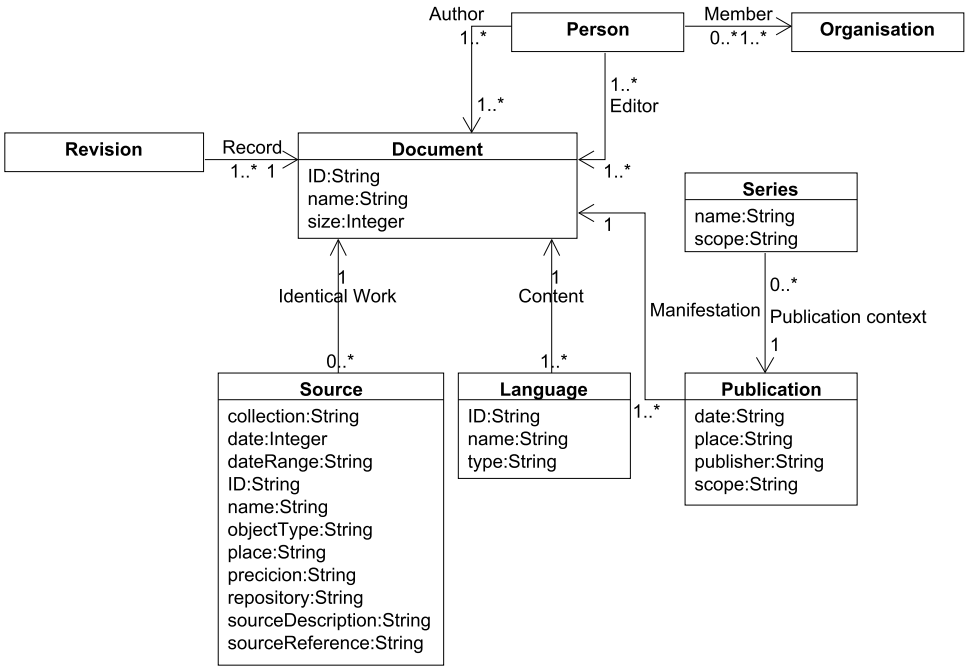

Abbildung 6.9: *Die Klasse **Document** mit ihren Attributen sowie weitere Relationen zu anderen Klassen mit ihren Attributen. Einem Objekt der Klasse **Document** werden mögliche weitere Quellen, die demselben Werk zugeschrieben sind, ein oder mehrere Sprachen sowie eine Veröffentlichung zugeordnet. Dazu werden jedem Objekt dieser Klasse noch Personen, die als Autor oder Herausgeber des Dokumentes auftreten, sowie ein oder mehrere Sprachen zugeordnet. Verkürzt, ohne Attribute dargestellt sind alle Klassen, die bereits vorgestellt wurden.*

Jedem Document werden beliebig viele Sprachen zugeordnet und können dann über das Attribut *ID* z. B. mit dem ISO Code[139] identifiziert werden. Im Metamodell ist bislang nur ein allgemeiner Fall *String* abgebildet. So kann für die Instanz

[139] Z. B. nach der International Organization for Standardization (ISO) 639 für Sprachen `http://www.iso.org/iso/home/standards/language_codes.htm` (besucht am 08.10.2016).

Alchymistische Practic[140] aus RIDGES angegeben werden, dass sie Deutsch und Latein enthält.

Die Klasse **Revision** bildet alle für ein Objekt der Klasse **Document** relevante Informationen ab, die durch Veränderungen von Objekten der Klasse **Annotation** entstanden sind, beispielsweise wenn Annotationen in einem Dokument entfernt, korrigiert oder hinzugefügt werden (vgl. Wiederverwendungsszenarien in Kapitel 3). Weiterhin können neue Dokumente in einer Version eines Korpus hinzugefügt werden, die dann wiederum Annotationen erhalten, ohne dass alle bereits vorhandenen Dokumente noch einmal einer Änderung unterliegen. So wird direkt ersichtlich, welche Teile eines Korpus in einer Revision von welchen Änderungen betroffen sind.

Die Klasse **Publication** umfasst deskriptive, strukturelle Metadaten. Ein Objekt der Klasse **Publication** wird mit Veröffentlichungsdatum, Veröffentlichungsort, Verlag und Umfang des Textes beschrieben (typische bibliographische Angaben). Wenn ein Objekt der Klasse **Publication** in einer Serie veröffentlicht ist, dann können jedem Objekt der Klasse **Publication** ein oder mehrere Objekte der Klasse **Series** als Publikationskontext zugeordnet werden.

Ein Objekt der Klasse **Document** repräsentiert eine historische Vorlage. Diese Vorlage liegt als Manifestation veröffentlicht vor (Publication). Mehrere Publication können auch einem Document zugeordnet werden. Eine Veröffentlichung kann beispielsweise eine Manifestation eines Werks abbilden. Wenn einem Objekt der Klasse **Document** weitere Manifestationen oder auch Expressionen (desselben Werks) als Quelle zugeordnet werden sollen, dann kann dies mit der Klasse **Source** abgebildet werden.

Nehmen wir die Beispiele *Otfrid* aus Abbildung 2.3 und *New Kreüterbuch* aus Abbildung 6.3. Die Instanz *Otfrid* der Klasse **Document** erhält weitere Metadaten, die über die Angaben hinausgehen, die dem Publication zugeordnet werden können, vgl. Abbildung 2.3. *New Kreüterbuch* ist ebenfalls eine Instanz der Klasse **Document** und erhält hingegen keine weiteren Angaben (vgl. Abbildung 6.10).

[140] Alchymistische Practic / Andreas Libavius (Frankfurt, 1603) [S. 4-19, 5504 dipl-Einheiten]

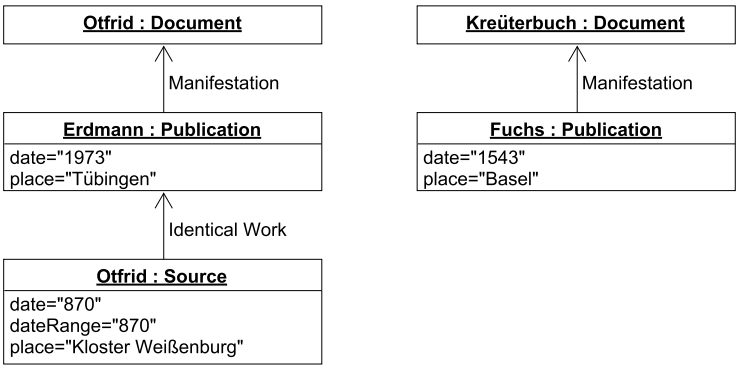

Abbildung 6.10: *Instanzenmodell für zwei Objekte der Klasse* ***Document*** *(Ausschnitt). Für die Instanz Otfrid werden eine Manifestation und eine weitere Quelle angegeben. Für die Instanz New Kreüterbuch ist nur eine Manifestation angegeben.*

Im MKM können mehrere Objekte der Klasse **Source** einem Objekt der Klasse **Document** zugeordnet werden. Es gilt die Bedingung, dass Objekte der Klasse **Source** zu demselben Werk wie das Objekt der Klasse **Document** zugeordnet werden können. Ein Objekt der Klasse **Source** kann ein publizierten Text oder ein unpublizierter Text (z. B. Expression) sein. Die Instanz der Klasse **Document** *Otfrid* in Abbildung 6.10 ist mit einer modernen Publikation von Oskar Erdmann (Erdmann : Publication)[141] und einer althochdeutschen Publikation von *Otfrid* als originalen Autor des Evangelienbuchs (Otfrid : Source) dokumentiert. Die Instanz *New Kreüterbuch* ist hingegen nur mit einer Publikation (Fuchs : Publication) dokumentiert. Somit versucht das MKM den Ansatz zur Beschreibung von Publikationen der Functional Requirements for Bibliographic Records (FRBR) zu integrieren.

Diese Metadaten, die durch die Document, durch die Publication sowie durch die Source beschrieben werden, sind von den im Korpus enthaltenen Transkriptionen oder Normalisierungen unabhängige Metadaten. So kann nicht automatisch davon ausgegangen werden, dass die Dokumentation einer Quelle wie im Beispiel von *Otfrid* darauf verweist, dass auf dieser Vorlage digitalisiert wurde und sich dies direkt in den im Dokument vorhandenen Annotationen widerspiegelt. Im Referenzkorpus Altdeutsch wurde beispielsweise die Edition von Erdmann als Vorlage für die

[141] Gemeint ist Erdmann (1973). Vgl. für einen kompletten Überblick über die Überlieferung von Otfrids Evanglienbuch `http://www.handschriftencensus.de/werke/1285` (besucht am 27.01.2017).

Digitalisierung genutzt. Diese wiederum bezieht sich aber auf eine weitere Quelle. Umgekehrt heißt dies nicht, wenn nur eine zusätzliche historische Manifestation für ein Dokument angegebenen ist, dass diese mit minimalem Interpretationsspielraum als eine Annotation repräsentiert sein muss. Jede Transkription kann wieder eigene Interpretationen der Vorlage besitzen und nach eigenen Richtlinien eingesetzt werden (vgl. Kapitel 2). Digitalisierungen repräsentieren eine historische Vorlage (Document) wiederum anders, wenn sie beispielsweise stark normierend eingreifen. Das jeweilige Exemplar kann auch einen Einfluss auf die Annotation haben, wenn beispielsweise in einem Exemplar Abschnitte nicht mehr lesbar sind.

Das Beziehungsgeflecht zwischen dem (historischen) Werk, seinen ggf. diversen Expressionen und Manifestationen, Exemplaren und den Objekten der Klasse **Document** sowie den Objekten der Klasse **Annotation** ist komplex und nicht immer direkt oder einheitlich aufeinander abbildbar (vgl. Abschnitt 2.7.1). Somit wird auch im MKM kein Versuch unternommen, dies direkt abzubilden zu wollen.

Die Metadaten, die in der Klasse **Document** und den dazu assoziierten Klassen modelliert werden, erfüllen die Handlung 1 (Description). Sie sind wesentlich für Handlung 3 (Retrieval), da ähnlich wie bei Bibliotheken die Suche und Auswahl von Korpora auf Grundlage ihrer enthaltenen Dokumente fußt. Weiterhin können Korpora durch diese Metadaten nach bibliographischen Kriterien katalogisiert werden, Handlung 2 (Management).

6.3.3 Die Klasse Corpus

Die Klasse **Corpus** existiert nicht unabhängig von der Klasse **Document**. Diese Klasse bildet alle deskriptiven und strukturellen Metadaten für Korpora ab.

Abbildung 6.11 zeigt die Klasse **Corpus** mit ihren Attributen sowie weiteren Assoziationen zu anderen Klassen und ihren Attributen. Verkürzt dargestellt sind alle Klassen, die bereits vorgestellt wurden.

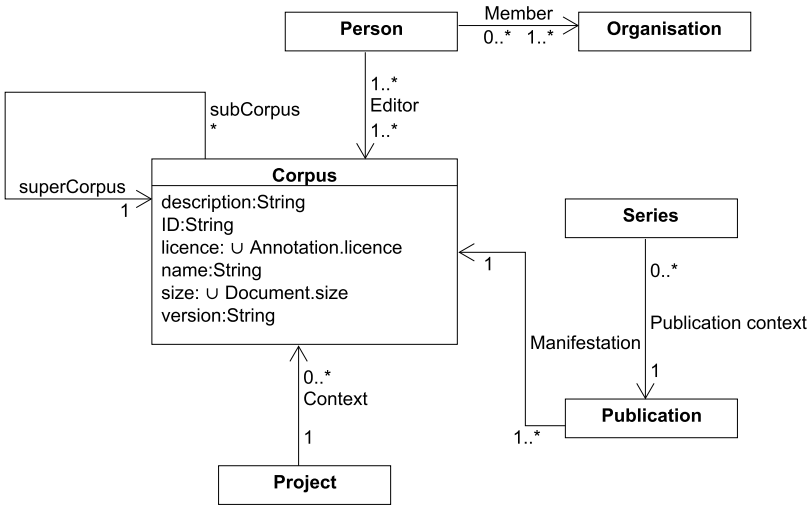

Abbildung 6.11: *Die Klasse* ***Corpus*** *mit ihren Attributen sowie weiteren Relationen zu anderen Klassen mit ihren Attributen. Einem Objekt der Klasse* ***Corpus*** *werden ein Projekt, Person, die als Herausgeber des Korpus auftreten, sowie mindestens ein Publication zugeordnet.*

In diesem Modell werden Beziehungen zwischen mehreren Objekten der Klasse **Corpus** so abgebildet: Jedes Objekt der Klasse **Corpus** kann als *Sub-Korpus* einem *Super-Korpus* zugeordnet werden.

Jedes Objekt der Klasse **Corpus** wird mit einem Namen beschrieben und wird keinen, einem oder mehreren Projekten zugeordnet. Weiterhin werden einem Corpus ein oder mehrere Person in der Rolle Herausgebers (*Editor*) zugeordnet. Damit sind das administrative Metadaten, die Informationen zum verantwortlichen Projekt und Herausgeber für das gesamte Korpus machen. Jedem Corpus werden ein oder mehrere Publication zugeordnet, das deskriptive und strukturelle Metadaten zum Veröffentlichungsdatum, -ort und -umfang des gesamten Korpus enthält. Durch ein oder mehrere Series können administrative und strukturelle Metadaten für das Korpus angegeben werden, wenn es in einer Reihe von Korpora veröffentlicht wird.

.Mit dem Attribut *description* kann eine Beschreibung des Korpus gegeben werden. Das Attribut *version* gibt die Version des Korpus an. Über das Attribut *licence* der Klasse **Corpus** wird die Lizenz des Korpus angegeben, die sich aus den Lizenzen der Annotationen ergeben. Auf dieselbe Weise wird in dem Attribut *size* die Größe des Korpus als Komposition aus mehreren Dokumenten und Annotationen angegeben.

Die Klasse **Corpus** fasst somit in erster Linie strukturell die Eigenschaften der enthaltenen Objekte der Klasse **Document** und indirekt der Klasse **Annotation** zusammen, die Handlung 2 (Management) und Handlung 1 (Description) ermöglichen. Weitere administrative Metadaten zur Veröffentlichung und zu Verantwortlichen ermöglichen die Handlung 4 (Authentifizierung) des Korpus. Deskriptive Metadaten zu Namen oder Sprachen des Korpus ermöglichen Handlung 3 (Retrieval).

6.3.4 Metamodell für Korpusmetadaten

Das Fundament des MKM bilden die drei Klassen **Annotation**, **Document** und **Corpus**. Das Metamodell beschreibt so aus einer technisch-abstrakten Perspektive ein Korpus als Menge seiner Dokumente und Annotationen. Abbildung 6.12 zeigt die Struktur des MKM mit allen seinen Klassen, ohne deren Attribute.

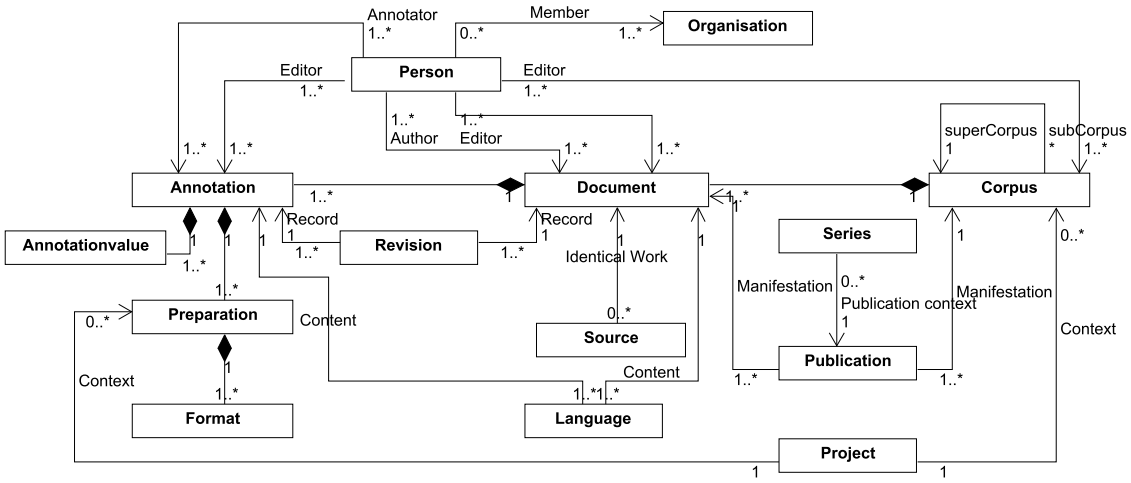

Abbildung 6.12: *Das Metamodell für Korpusmetadaten (MKM). Darstellung mit allen Klassen und Relationen zwischen den Klassen, ohne Attribute.*

Die Klasse **Person** wird in verschiedenen Rollen allen drei Klassen zugeordnet. So wird abgebildet, dass Annotationen, Dokumente und Korpora durch verschiedenen Personen erstellt, verarbeitet oder veröffentlicht werden (können). Häufig sind die Autoren eines Objektes der Klasse **Document** keine Autoren eines Objektes der Klasse **Annotation**. Gerade bei historischen Dokumenten können dies nicht dieselben Personen sein. Annotatoren einer Annotation müssen ebenfalls nicht zwangsläufig auch die Herausgeber einer Korpus sein und umgekehrt. Eine zusätzliche administra-

tive Information wird über die Angabe des Kontextes (Project) für jede Annotation und für das Korpus angegeben.

Jeweils einem Objekt der Klasse **Corpus** und der Klasse **Document** wird mindestens ein Objekt der Klasse **Publication** zugeordnet. Typischerweise sind Dokumente in irgendeiner Form vor der Korpuserstellung veröffentlicht worden. Annotationen können nicht unabhängig von dem Korpus, in dem sie enthalten sind, veröffentlicht, sehr wohl aber verändert werden.

Diese drei zentralen Klassen werden aus einer produktorientierten Perspektive mit Metadaten beschrieben; das Korpus wird als Produkt aus verschiedenen Arbeitsschritten definiert. Die Eigenschaften, die Annotationen zu einem bestimmten Zeitpunkt besitzen, sowie die konkreten Bearbeitungsschritte, die für deren Erstellung und Nachvollziehbarkeit notwendig sind, stehen im Fokus. So werden neben den vorhandenen Annotationswerten und deren Erklärungen die einzelnen abgeschlossenen Bearbeitungsschritte in einem oder mehreren Formaten eines Korpus mit Metadaten beschrieben. Damit werden die Formate neben den Annotationen referenziert, da Annotationen in einem oder mehreren Formaten vorhanden sein können, aber nicht alle Formate alle Annotationen abbilden müssen. Nicht alle Texte eines Korpus müssen jeweils die gleiche Menge an Annotationen enthalten. So werden Annotationen und Dokumente einzeln auch in den Metadaten referenziert. Ein Korpus als Komposition aus mindestens einem Dokument wird ebenfalls referenziert. So kann beispielsweise eine Metadatensuche auf die klassenspezifischen Metadaten aufbauen. Gesucht werden kann dann nach Eigenschaften bestimmter Annotationen, Formaten, Dokumente oder Korpora, vgl. Handlung 3 (Retrieval).

Durch die Modellierung eines Korpus als Komposition aus Dokumenten und Annotationen wird es ebenfalls möglich, strukturelle Eigenschaften des Korpus im Detail durch Metadaten zu beschreiben. Ein Katalogisieren der im Korpus enthaltenen Dokumente sowie der Annotationen ist so möglich. Korpora können dann auf der Korpusebene, der Dokumentebene und der Annotationsebene eingeordnet, verglichen und sortiert werden, vgl. Handlung 2 (Management).

Die verschiedenen Klassen erhalten Referenzen, wie **Document** und **Source**, **Annotation**, **Language** und **Corpus**, über die sie über ihre eigentliche Instanziierung hinweg identifiziert werden können. Bislang ist der Werttyp des Attributs ID als *String* festgelegt. Im MKM ist damit nur der allgemeine Fall abgebildet. Die Wahl der ID ist somit frei und für jede spezifische Anwendung, für jedes Tool, muss eine bestimmte ID-Lösung festgelegt werden. [142]

[142] In einer konkreten Anwendung des MKM wie z.B. in einem Repositorium muss jeweils interpre-

Die einzelnen Wiederverwendungsszenarien aus Kapitel 3 können einzelne Annotationswerte oder Annotationen, Bearbeitungsschritte, Formate, Dokumente und Sprachen oder Personen betreffen. Deren Änderung können wiederum die gesamte Korpusarchitektur und -veröffentlichung beeinflussen. Die Metadaten, die das MKM modelliert, ändern sich je nach Szenario für ein Korpus pro Version entweder nur punktuell oder umfangreicher, wenn beispielsweise mehrere Wiederverwendungsszenarien auf ein Korpus angewandt werden. Die umfangreichen Informationen über die Annotationen, die Erstellung und die Architektur des Korpus müssen dann Dritten vorliegen, um ein oder mehrere Wiederverwendungsszenarien durchführen zu können. Darauf folgt, dass Änderungen am Korpus durch die getätigten Wiederverwendungsszenarien von den verschiedenen Akteurinnen und Akteuren in der Korpusdokumentation festgehalten werden. So kann ein Dokumentationszyklus entstehen, der das Korpus nach jeder erneuten Wiederverwendung einheitlich als Komposition aus Dokumenten (und diese aus Annotationen) und als Produkt verschiedener Bearbeitungsschritte beschreibt. Jede weitere Veränderung – Wiederverwendungsszenario – wird auf dieselbe tief strukturierte, einheitliche Weise mit Bezug zum Forschungsdatenzyklus und zum Forschungsprozess in einer Korpusdokumentation nach dem MKM festgehalten.

Die Eigenschaften der Objekte der Klassen **Annotation**, **Document** und **Corpus** und die Klassen **Person**, **Project**, **Publication** sind in dem MKM abgebildet und einheitlich, technisch-abstrakt, produktorientiert und umfangreich durch administrative, strukturelle, deskriptive und technische Metadaten beschrieben, vgl. Handlung 1 (Description).

Wie bei einem Haus, für das ein An- oder Umbau geplant ist, müssen Informationen über vorhandene Räume, deren Funktionen, und die Eigenschaft der Wände (tragend oder und nichttragend) vorliegen. In ähnlicher Weise enthält ein Dokument eine Menge an Annotationen, wie ein Raum eine Menge an Wänden enthält. Dieses Aggregat an Annotationen teilt sich eine Menge an Metadaten – die der Dokumente. Die Menge an Wänden, die einen Raum bilden, erhält dann eine Funktion, z. B. *Küche* oder *Flur*. Ein Haus besteht dann aus einem oder mehreren Räumen mit unterschiedlichen Funktionen, wie ein Korpus aus einem oder mehreren Dokumenten besteht. Wenn Räume hinzukommen, müssen Wände gebaut oder geändert werden. Gleiches gilt für Dokumente und Annotationen. Dabei ist mit Blick auf die Korpusarchitektur wichtig, welche Annotationen eine eigenständige Segmentierung besitzen, also tragende Wände eines Hauses darstellen. Gleiches gilt, wenn ein Hausbau nach-

tiert werden, welche konkrete IDs (z. B. URL zu einer ISO-Cat-Referenz) vorliegen.

vollzogen oder repliziert wird, um bei der Haus-Metapher zu bleiben. Damit wird für jede Version eines Korpus eine Dokumentation erstellt, die damit den Aufbau und Inhalt des Korpus abbildet. Durch alle Revision gibt es die Möglichkeit, für alle Annotation und alle Document Veränderungen jeder Version aufzulisten.

Mit diesem Abschnitt wird gezeigt, dass die Eigenschaften eines Korpus in Form von Metadaten im MKM so modelliert werden, dass auf deren Basis Handlungen wie Beschreibung, Katalogisieren, Suche und Authentifizieren ermöglicht werden, um dann in einem nächsten Schritt ein oder mehrere Wiederverwendungsszenarien (z. B. Erweiterung des Korpus) durchführen zu können. Die Struktur des Bauplans – die Korpusdokumentation – ist dabei einheitlich und technisch-abstrakt, um eine ersteller- und überfachliche Erschließung von Korpora über deren Metadaten zu ermöglichen.

Das MKM liegt mit dieser Arbeit in seiner ersten Version vor. Es sollte aufgrund seiner Architektur erweitern werden können. Für ein konkretes Anwendungsszenario muss das MKM bzw. die Objekte der jeweiligen Klassen in einer festen, konkreten Datenstruktur vorliegen. In Kapitel 7 wird die TEI als Metadatenmodell für eine Realisierung des MKM verwendet.

7 Realisierung des Metamodells für Korpusmetadaten

In einem nächsten Schritt ist eine Realisierung des MKM in verschiedene Metadatenformate wie DCMES (ISO 2014), TEI (TEI Consortium 2015) oder CMDI (Broeder et al. 2010) möglich. Idealerweise können die Instanzen des MKM in mehreren Formaten abgebildet werden. Wie Kapitel 5 gezeigt hat, besitzt der Ansatz der TEI eine Modellierungssprache (ODD) und ein Modell von Dokumenten, das durch komplexe und umfangreiche Guidelines abgebildet wird. Mit der ODD kann eine Anpassung der TEI nachvollziehbar und transparent modelliert bzw. spezifiziert werden. Weiterhin wird der Inhaltsstandard der FRBR für TEI-Dokumente umsetzt. Darauf kann daher für eine Realisierung des MKM aufgebaut werden. In diesem Abschnitt wird ein Vorschlag erarbeitet, wie mit Hilfe der TEI die Korpusmetadaten des MKM realisiert werden können.

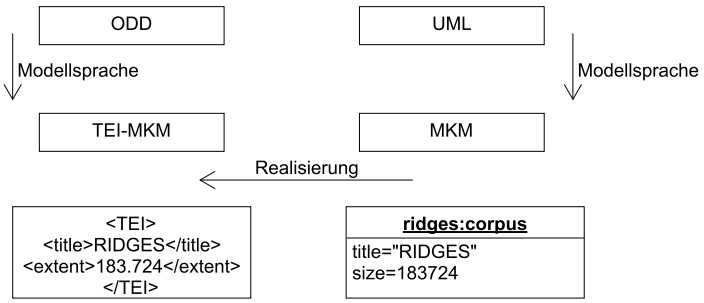

Abbildung 7.1: *Realisierung des MKM in TEI. Das Verhältnis von UML als Modellsprache für das MKM ist vergleichbar mit der ODD als Modellsprache für TEI(-MKM). Mit der TEI-Spezifikation werden die Objekte der Klassen des MKM in TEI realisiert.*

Die Instanziierungen des MKM sind jeweils Modelle von Korpusmetadaten von historischen Korpora. Diese Metadatenmodelle müssen in einem weiteren Schritt realisiert werden, um sie für Anwendungen nutzbar zu machen. Die TEI wird durch

eine Spezifikation für die Metadatenmodelle (Instanzen des MKM) angepasst. Durch diese Spezifikation werden auch Schemata erstellt, wodurch die Korpusmetadaten in einem validen, TEI-konformen Format gespeichert werden können.

Das Zusammenspiel dieser Komponenten funktioniert so: Das MKM gibt z. B. für alle Objekte der Klasse **Document** ein Attribut *date* an, womit das Erscheinungsjahr eines Dokumentes angegeben werden kann. Das Dokument eines Korpus wird mit einer Jahresangabe beschrieben. Dieses Metadatum muss in TEI realisiert werden. Für die Objekte der Klasse **Document** wird eine TEI-Spezifikation erstellt. Diese Spezifikation wählt das TEI-Element `<date>` [143] für diese Information aus und bettet es in eine TEI-konforme Struktur. Die Korpusmetadaten können dann in einer TEI-XML-Datei als `<date>Jahr</date>` vorliegen.

Das TEI-Modell als Metadatenmodell und deren Format als seine Instanziierung zu verwenden, ist ein neuer Ansatz, da der Anwendungsbereich der TEI originär in der Kodierung von Dokumenten liegt (Abschnitt 5.5), deren Bestandteil auch dessen Metadaten sind.[144] Der Ansatz dieser Arbeit stützt sich auf die Stärken der TEI hinsichtlich einer flexiblen Modellierung und die umfassende Erfahrung der einzelnen Fachgebiete, die die TEI stetig weiterentwickeln. Vgl. Abschnitt 5.5 für weitere Informationen zur TEI.

Der nächste Abschnitt beschreibt kurz, wie eine TEI-Spezifikation funktioniert, bevor die drei in Rahmen dieser Arbeit erstellten Spezifikationen vorgestellt werden.

7.1 TEI-Spezifikationsdokument ODD

Die Spezifikation der TEI erfolgt mit dem TEI eigenen Modell, der One Document does it all (ODD) (Burnard und Rahtz 2004). Das ODD-Spezifikationsdokument ermöglicht es, Module, Elemente und Attribute der TEI-Guidelines zusammenzustellen, neu zu strukturieren und zu spezifizieren. Grundlage eines jeden Spezifikationsdokumentes kann entweder das TEI-ALL-Schema[145] sein, das dann mittels der

[143] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-date.html` (besucht am 30.10.2016).

[144] Mit dem Institute for Corpus Linguistics and Text Technology (ICLTT) Metadaten Initiative wurde ein Ansatz TEI-Header für das Austrian Academy Corpus, diskutiert. Die enthaltenen Texte liegen in als TEI-konforme Dokumente vor und werden als solche hauptsächlich auf der Dokumentebene beschrieben (vgl. Budin et al. 2012), so dass der Ansatz nicht über die TEI-Welt hinausgeht.

[145] `http://www.tei-c.org/release/xml/tei/custom/schema/relaxng/tei_all.rng` (besucht am 06.07.2016).

ODD eingeschränkt wird, oder das TEI-BARE-Schema[146], das mittels der ODD aufgebaut wird. So ist es möglich, ein eigenes Subset (Untermenge) und eine angepasste Modulstruktur zu entwickeln:

> In the ODD model, a markup scheme is defined by a schema, which may subsequently be instantiated using the concrete syntax of an XML DTD, a RelaxNG schema, or a W3C Schema. An ODD schema consists of references to a number of discrete modules, combined more or less as required; the distinction between 'base' and 'additional' tagsets in earlier versions of the Guidelines has not been carried forward into P5. The modules referenced comprise declarations of particular elements and attributes, which can be combined with further declarations given explicitly in the schema for customization purposes.(Burnard und Rahtz 2004: 2)

Das ODD-Spezifikationsdokument ist für eine konkrete Anwendung eine Art Multifunktionsdokument. Es enthält eine Beschreibung des Anwendungsziels und des Anwendungskontextes sowie Beispiele für den Gebrauch. Das Dokument enthält ein formale Angabe der Komponenten des *TEI Abstract Model*: Elemente und Attribute, Module und Klassen.[147] Aus der ODD ist es weiterhin möglich, formale Schemata wie z. B. Document Type Definition (DTD)[148] oder REgular LAnguage for XML Next Generation (RELAX NG)[149] zu generieren.

In dieser Arbeit wird damit eine anwendungsorientierte Anpassung der TEI P5 (P5 steht für Version 5) gemacht. Hervorgehoben werden soll, dass die hier vorgeschlagene Modellierung ein reines Subset der TEI darstellt. Es wurden keine neuen Elemente, Attribute und Module definiert und in die bestehenden Strukturen integriert. Das Ziel war es, auf die bestehende Modelle der TEI aufzubauen, so dass die TEI-XML-Struktur aus `teiHeader` und `body` (bzw. `text`) weiterhin bestehen bleibt und eine größtmögliche Chance auf Interoperabilität mit anderen TEI-Lösungen besteht (vgl. Abschnitt 5.5).

Jede zentrale Klasse des MKM erhält eine eigene Spezifikation durch die ODD. Eine wesentliche Änderung der Perspektive wird für die hier entwickelten TEI-Spezifikationen vorgenommen: Der Bezug des `teiHeader` wird von dem Dokument,

[146] `http://www.tei-c.org/Guidelines/Customization/index.xml?style=raw` (besucht am 06.06.2016).
[147] Vgl. `http://wiki.tei-c.org/index.php/ODD` (besucht am 15.01.2017).
[148] `http://www.w3schools.com/xml/xml_dtd.asp` (besucht am 27.01.2017).
[149] `http://relaxng.org/` (besucht am 15.01.2017).

das in derselben TEI-konformen Datei vorliegt, gelöst (Abbildung 7.2).

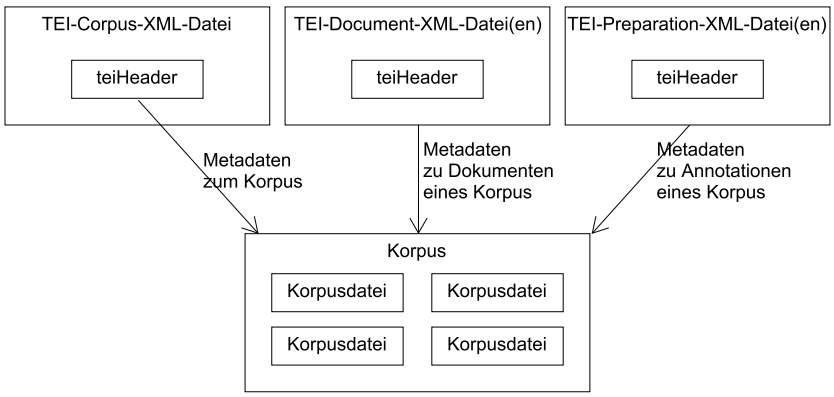

Abbildung 7.2: *TEI-Dateien, deren* `teiHeader` *jeweils die Metadaten eines Objektes einer Hauptklasse sowie Objekte anderer assoziierter Klassen trägt, dessen* `body` *(oder* `text`*) aber leer bleibt.*

Die jeweiligen Realisierungen der Korpusmetadaten erfolgt in TEI-konformen Dateien, genauer im `teiHeader`. Da die Korpora selbst in eigenen Formaten (dann konkret Dateien) vorliegen, werden die zu beschreibenden Objekte nicht selbst in TEI-konformen Dateien integriert. Das eigentliche Dokument der TEI-konformen Datei – `body` oder `text` – ist leer, was Abbildung 7.2 illustriert. Ein Korpus wird unabhängig von den `teiHeader` als Sammlung von Korpusdateien dargestellt. Für jedes Objekt der Klassen **Annotation**, **Document** und **Corpus** wird eine TEI-Spezifikation und für deren Realisierung jeweils eine TEI-XML-Datei verwendet. Die TEI wird damit als ein Modell für Metadaten genutzt. Abbildung 7.3 zeigt ein Beispiel für TEI-Metadaten von RIDGES.

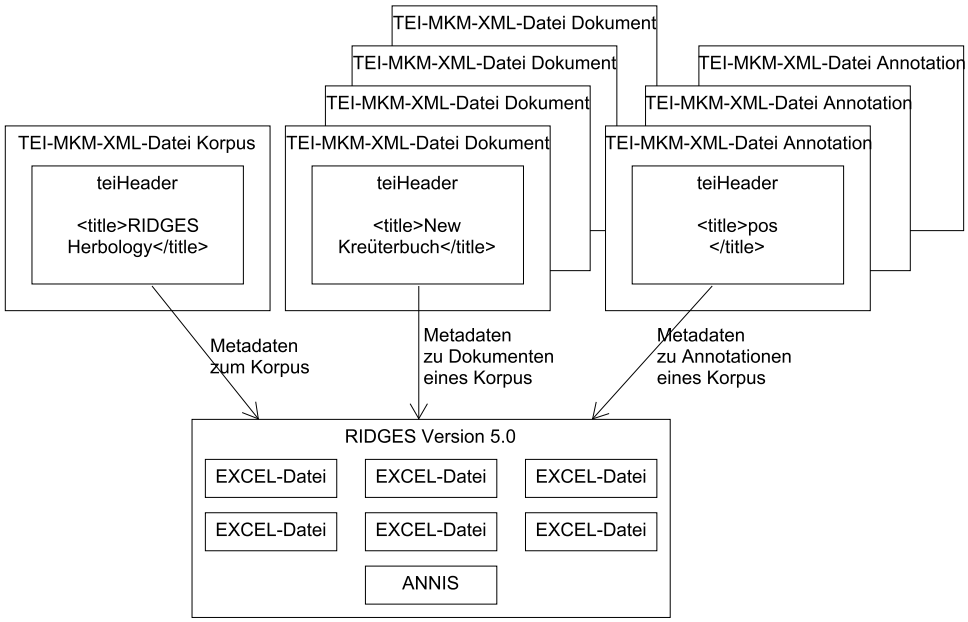

Abbildung 7.3: *TEI-Metadaten am Beispiel von RIDGES. Die Objekte der Klassen des MKM repräsentieren eine bestimmte Auswahl an Metadaten des RIDGES-Korpus. Bezogen auf die drei Hauptklassen und ihren Relationen zu anderen Klassen im MKM gibt es drei TEI-Spezifikationen. Metadaten zum Objekt der Klasse* ***Corpus*** *können in einer TEI-MKM-XML-Datei für Korpora realisiert werden, Metadaten zu Objekten der Klasse* ***Document*** *jeweils in einer TEI-MKM-XML-Datei für Dokumente und Metadaten zu Objekten der Klasse* ***Annotation*** *jeweils in einer TEI-MKM-XML-Datei für Annotationen.*

Das RIDGES-Korpus in der Version 5.0 liegt in einer ANNIS-Datei und in mehreren EXCEL-Dateien vor. Eine Auswahl an Eigenschaften (Metadaten), die das RIDGES-Korpus an sich besitzt, wird als Objekt der jeweiligen Klassen des MKM repräsentiert. Diese Metadaten werden dann jeweils passend in die verschiedenen TEI-Spezifikationen realisiert. So wird beispielsweise für das Objekt der Klasse **Corpus** der Name angegeben, genauso wie für alle Objekte der Klasse **Document** und **Annotation**. In der TEI-MKM-XML-Datei für das RIDGES-Korpus wird im `teiHeader` mit dem Element `<title>` dieses Metadatum realisiert. Für jedes Objekt der Klasse **Document** wird eine TEI-MKM-XML-Datei für Dokumente angelegt, in der dann jeweils der Name des Dokuments mit dem Element `<title>` angegeben wird. Gleiches gilt für die Objekte der Klasse **Annotation** und die TEI-MKM-XML-Datei(en) für Annotationen. Diese TEI-MKM-XML-Dateien liegen unabhängig von den Kor-

pusdateien vor. Damit wird die TEI mit Hilfe des MKM als Metadatenmodell und Datenstrukturstandard genutzt.

Im Folgenden werden jeweils die Realisierung allgemein und einige für den Anwendungsfall wichtige kleinere Strukturen vorgestellt. Alle Spezifikationsdokumente inklusive einer Dokumentation sind unter CC-BY 4.0 Licence[150] frei verfügbar (Odebrecht 2017).

7.1.1 Spezifikation für die Klasse Annotation

Für die Objekte der Klasse **Annotation** werden die TEI-Module `tei`, `textstructure`, `core`, `header` und `namesdates` in der ODD-Spezifikation genutzt.[151]

```
<TEI>
<schemaSpec ident="teiODD_LAUDATIOPreparation_S7.1">
    <moduleRef key="core" include="author editor date list item label p
ref title"/>
    <moduleRef key="tei"/>
    <moduleRef key="header"
        include="appInfo application authority availability change
correction editorialDecl encodingDesc extent fileDesc idno langUsage
language namespace normalization profileDesc projectDesc publicationStmt
revisionDesc segmentation sourceDesc tagUsage tagsDecl teiHeader
titleStmt"/>
    <moduleRef key="textstructure" include="TEI text"/>
    <moduleRef key="namesdates" include="affiliation persName forename
surname orgName"/>
    <!--  …  -->
</schemaSpec>
</TEI>
```

Abbildung 7.4: *Ausschnitt aus der TEI-Spezifikation für die Objekte der Klasse **Annotation** mit allen verwendeten Modulen und alle hinzugefügten Elementen.*

Die Anpassung dieser Module basiert hauptsächlich auf zwei Prinzipien: der Einschränkung der nutzbaren Elemente und deren Attribute sowie der Hinzufügung einiger weniger Elemente. Das Element `<moduleRef>` in der ODD gibt an, welche Module und welche Elemente oder Elementgruppen in welchen Modulen in der **Annotation**-TEI-Spezifikation genutzt werden (Abbildung 7.4).

[150] `https://creativecommons.org/licenses/by/4.0/deed.de` (besucht am 02.02.2017).
[151] Für eine Liste der Module vgl. `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ST.html#STMA` (besucht am 15.01.2017).

Die eingesetzten Module werden jeweils mit dem Element `<moduleRef>`[152] aufgelistet, dessen Attribut `@include` eine Liste aller Elemente angibt, die aus anderen Modulen kopiert und in das zu definierende Schema integriert werden. Keines der verwendeten Module wurde in seiner Struktur grundlegend verändert, es werden nur benötigte Elemente hinzugefügt. Einige Elemente wiederum werden hinsichtlich ihrer Attribute verändert. Beispielsweise werden nicht benötigte Attribute entfernt oder Attribute erhalten offenen oder geschlossene Wertelisten.

```
<TEI>

        <elementSpec ident="segmentation" module="header"
mode="change">
          <attList>
            <attDef ident="ana" mode="delete"/>
            <attDef ident="change" mode="delete"/>
            <attDef ident="copyOf" mode="delete"/>
            <attDef ident="corresp" usage="opt" mode="add"/>
      <!-- ... -->
            <attDef ident="sameAs" mode="delete"/>
            <attDef ident="select" mode="delete"/>
            <attDef ident="style" mode="change" usage="rec">
              <valList type="closed" mode="replace">
                <valItem ident="Dependent"/>
                <valItem ident="Independent"/>
              </valList>
            </attDef>
            <attDef ident="synch" mode="delete"/>
            <attDef ident="xml:base" mode="delete"/>
            <attDef ident="xml:id" mode="delete"/>
            <attDef ident="xml:lang" mode="delete"/>
            <attDef ident="xml:space" mode="delete"/>
          </attList>
        </elementSpec>
      </schemaSpec>

</TEI>
```

Abbildung 7.5: *Beispiel für eine Elementspezifikation des* `<segmentation>` *für die Objekte der Klasse* ***Annotation****. Viele Attribute von* `<segmentation>` *werden gelöscht, das* `@style` *wird verändert und das* `@corresp` *wird hinzugefügt.*

Neben den Modulspezifikationen können also ebenfalls einzelne Elemente durch die Anpassung ihrer Attribute spezifiziert werden. Diese Elementespezifikationen werden mit `<elementSpec>` und `<attList>` (für die Attribute der Elemente) in der ODD umgesetzt. Ein Beispiel für eine Elementenspezifikation des Elementes `<segmentation>`[153] aus dem Module `header` zeigt Abbildung 7.5.

[152] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-moduleRef.html` (besucht am 17.10.2016).

[153] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-segmentation.html`

Mehrere nicht benötigte Attribute des Elementes `<segmentation>` werden in dieser Spezifikation gelöscht – `mode="delete"` –, damit die Varianz der Realisierung des Elementes sehr stark eingeschränkt wird. Weiterhin kann der Gebrauch der Attribute von Elementen, wie in diesem Fall `<segmentation>`, über das Attribut `@usage` beispielsweise als erforderlich – rec – oder optional – opt – festgesetzt werden.[154] So werden in diesem Beispiel das Attribut `@corresp` als optional und das Attribut `@style` als erforderlich angegeben. Darüber hinaus können die Werte der Attribute eines Elementes `<segmentation>` in offenen oder geschlossenen Listen angegeben werden, in diesem Beispiel gibt es zwei Werte in einer geschlossenen Liste für das Attribut `@style`, nämlich *Dependent* und *Independent*. Diese Angaben sind beispielsweise elementar, um die Korpusstruktur mit ihren Annotationen sowie eigenständigen und abhängigen Tokenisierungen zu dokumentieren (Abschnitt 6.3.1).

Die Objekte der Klasse **Annotation** werden in einem `teiHeader` realisiert, der die strukturellen, administrativen, deskriptiven und technischen Metadaten einer Annotation eines Korpus enthält. Abbildung 7.6 zeigt die grobe Struktur des `teiHeader` ohne die Anzeige der tieferen XML-Struktur oder einzelner Werte. Ein typischer TEI-konformer `teiHeader` besteht aus den Elementen `<fileDesc>` mit `<titleStmt>`, `<publicationStmt>`, `<sourceDesc>`, `<encodingDesc>` und `<revisionDesc>`. Die `<fileDesc>` „enthält die detaillierte bibliografische Beschreibung einer elektronischen Datei“[155], die hier auf eine Annotation bezogen wird und ebenfalls bibliographische Beschreibungen erhalten kann. Das Element `<encodingDesc>` „dokumentiert das Verhältnis zwischen dem elektronischen Text und seiner Quelle oder den Quellen, von der oder von denen er abstammt“[156]. Zum Zweck der Darstellung des MKM in TEI wird `<encodingDesc>` nun so uminterpretiert, dass sie das Verhältnis zwischen Annotation und Umsetzung beschreibt. Die `<revisionDesc>` „enthält alle Revisionsschritte, die an der Datei vorgenommen wurden. “[157]. Auf die Klasse der **Annotation** uminterpretiert, enthält `<revisionDesc>` alle Revisionsschritte, die an einer Annotation durchgeführt worden sind.

Jedes Objekt der Klasse **Annotation** wird dann in einer TEI-konformen Datei

(besucht am 15.01.2017).

[154] Die Verwendung kann dazu analog auch als empfohlen, optional oder verbindlich ausgewiesen werden.`http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-attDef.html` (besucht am 17.10.2016).

[155] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-fileDesc.html` (besucht am 19.10.2016).

[156] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-encodingDesc.html` (besucht am 19.10.2016).

[157] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-revisionDesc.html` (besucht am 19.10.2016).

realisiert und die Metadaten zum Namen, den Herausgebern und den Annotatoren sowie der Veröffentlichung und Revisionsgeschichte werden umgesetzt (vgl. Abbildung 7.1). In Abbildung 7.6 wird ein Beispiel einer Realisierung für ein Objekt der Klasse **Annotation**.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader type="PreparationHeader">
    <fileDesc>
      <titleStmt>
        <title>...</title>
        <editor>...</editor>
        <author>...</author>
      </titleStmt>
      <extent>...</extent>
      <publicationStmt>...</publicationStmt>
      <sourceDesc>...</sourceDesc>
    </fileDesc>
    <encodingDesc>...</encodingDesc>
    <revisionDesc>...</revisionDesc>
  </teiHeader>
  <text/>
</TEI>
```

Abbildung 7.6: *Beispiel für Realisierung der TEI-Spezifikation für ein Objekt der Klasse **Annotation**.*

Die Abbildung 7.6 zeigt eine (leere) `teiHeader`-Struktur, in der die jeweiligen Metadaten einer Annotationsebene eingefügt werden können.

Eine wesentliche Elementgruppe für die Angabe von technischen Metadaten zu den produktorientierten Erstellungsschritten einer Annotation enthält `<encodingDesc>`. Die Abbildung 7.7 zeigt ein gekürztes Beispiel einer solchen `<encodingDesc>` für die Annotation *komp* – Annotation von Komposita – in RIDGES.

```
<encodingDesc n="1" style="SpanAnnotation">
          <appInfo n="1" style="Manual">
            <application ident="EXCEL" style="xls"
version="14.06123.5001" type="Morphological" subtype="NA">
              <label>MS Excel 2010.</label>
              <p>Span annotation with key 'komp' of the
all documents of RIDGES Herbology 4.1. Manually annotated
by annotators. Consistency checks and further changes such
as changes in tokenization are applied in this preparation
step. For further information see
http://korpling.german.hu-
berlin.de/ridges/documentation_v4.1_en.html and
http://korpling.german.huberlin.de/ridges/download/v4.1/co
nversion_to_annis.pdf.</p>
            </application>
          </appInfo>
          <editorialDecl>
            <segmentation style="Dependent"
corresp="dipl">
              <p>Segmentation depends on 'dipl'.</p>
            </segmentation>
            <normalization method="None">
              <p>NA</p>
            </normalization>
            <correction status="low" method="manual">
              <p>Manually checked by annotators.</p>
            </correction>
            <p>NA</p>
          </editorialDecl>
          <projectDesc>
            <p>
              <ref target="http://korpling.german.hu-
berlin.de/ridges/" />
              The RIDGES project (Register in Diachronic
German Science) is an investigation into the development
of the German scientific language in the early modern and
modern periods, ranging from the mid 16th to the late 19th
century. The LAUDATIO project (http://www.laudatio-
repository.org) hosts and curates the RIDGES Herbology
corpus in cooperation with the RIDGES project.
            </p>
          </projectDesc>
        </encodingDesc>
```

Abbildung 7.7: *Beispiel für die Angabe von technischen Metadaten zu einem Bearbeitungsschritt eines Objektes der Klasse* ***Annotation****. Jeder Bearbeitungsschritt der Annotationsebene komp aus RIDGES Version 5.0 wird mit verschiedenen Metadaten durch einen Abschnitt* `encodingDesc` *beschrieben.*

Ein Objekt der Klasse **Annotation** existiert nicht unabhängig von den Objekten der Klasse **Preparation** (Abschnitt 6.3.1). Letztere sind mit dem `encodingDesc` reali-

siert. Jede Annotation kann mit einem oder mehreren Schritten – `encodingDesc@n` – beschrieben werden. Demnach können in der TEI-XML-Datei mehrere Erstellungsschritte einer Annotation mit `<encodingDesc>` angegeben werden. In dem Fall von *komp* wird in einem ersten Bearbeitungsschritt angegeben, um welche Art des Schritts es sich handelt – `@style="SpanAnnotation"`. Danach werden Metadaten über die verwendete Anwendung – `<appInfo>`[158] – angegeben.

Abbildung 7.7 zeigt, dass im Rahmen des Projektes RIDGES (`<projectDesc>`) die Annotation *komp* als Spannenannotation (`encodingDesc/@style`) im EXCEL-Format (`application/@ident`) manuell (`appInfo/@style`) annotiert wurde. Die Annotation *komp* besitzt eine abhängige Segmentierung (`segmentation/@style`) und wurde manuelle überprüft (`correction/@method`). Weiterhin erhält man auch eine freie Referenz auf die Annotationguidelines des Korpus (`<p>`).

Die TEI-MKM-Spezifikation für Annotationen enthält Metadaten der Objekte der Klassen **Annotation**, **Preparation**, **Format**, **Revision** sowie administrative Metadaten zu Verantwortlichen, die als Objekte der Klassen **Person** und **Project** beschrieben werden können. Die deskriptiven Metadaten zu den enthaltenen Werten einer Annotation werden in der Spezifikation zu Objekten Klasse **Corpus** realisiert und über eine ID referenziert, die in dieser Spezifikation zur Klasse **Annotation** angegeben wird (Abschnitt 7.1.3).

Der Aufbau der TEI-XML-Datei im Bereich des `teiHeader` wird mit dieser hier vorgestellten ODD-Spezifikation definiert. In dem typischen Aufbau einer TEI-XML-Datei wird der eigentliche Inhalt, der in dem `teiHeader` beschrieben wird, aufgeführt (vgl. Abschnitt 5.5). Da der hier vorgeschlagene Ansatz die TEI als eine Metadatenmodell nutzt, bleibt der `<body>` bzw. `<text>` leer, in dem die eigentliche Inhalte eines Dokumentes untergebracht werden. Eine mögliche weitere Entwicklung, die über den Rahmen dieser Arbeit hinaus geht, wäre es, auch Inhalte zu den passenden Metadaten in dieser Spezifikation zu erlauben und in der konkreten TEI-XML-Datei umzusetzen (Kapitel 8).

7.1.2 Spezifikation für die Klasse Document

Für die Objekte der Klasse **Document** werden die TEI-Module `tei`, `textstructure`, `core`, `header`, `tagdocs` und `namesdates` in der ODD-Spezifikation genutzt. Die Anpassung dieser Module basiert wie bei der Spezifikation für die Objekte der Klasse **Annotation** auf zwei Prinzipien: die Einschränkung der nutzbaren Elemente und

[158] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-appInfo.html` (besucht am 18.10.2016).

deren Attribute sowie der Hinzufügung einiger weniger Elemente. Das Element `<moduleRef>` gibt an, welche Module und welche Elemente oder Elementgruppen in welchen Modulen in der **Document**-TEI-Spezifikation genutzt werden, wie der folgende Auszug aus der ODD in Abbildung 7.8 zeigt.

```
<TEI>
<schemaSpec ident="teiODD_LAUDATIODocument_S7.1">
    <moduleRef key="core"
        include="author biblScope editor date p pubPlace publisher ref schemaSpec title"/>
    <moduleRef key="tei"/>
    <moduleRef key="header"
        include="change encodingDesc extent fileDesc idno langUsage language
        profileDesc publicationStmt revisionDesc seriesStmt sourceDesc teiHeader titleStmt"/>
    <moduleRef key="textstructure" include="TEI text"/>
    <moduleRef key="tagdocs" include="schemaSpec elementSpec valList valItem"/>
    <moduleRef key="namesdates" include="persName forename surname orgName"/>
    <!-- ... -->
</schemaSpec>
</TEI>
```

Abbildung 7.8: *Ausschnitt aus der TEI-Spezifikation für die Objekte der Klasse **Document**. Hier werden alle verwendeten Module der TEI sowie alle zusätzlich hinzugefügten Elemente aufgelistet.*

Die eingesetzten Module werden jeweils mit dem Element `<moduelRef>`[159] aufgelistet, dessen Attribut `@include` eine Liste aller Elemente angibt, die aus anderen Modulen kopiert und in das zu definierende Schema integriert werden. Keines der verwendeten Module wurde in seiner Struktur grundlegend verändert, es werden nur benötigte Elemente hinzugefügt. Einige Elemente wiederum werden hinsichtlich ihrer Attribute verändert. Beispielsweise werden nicht benötigte Attribute entfernt oder Attribute erhalten offenen oder geschlossene Wertelisten.

Neben den Modulspezifikationen können ebenfalls einzelne Elemente durch die Anpassung ihrer Attribute spezifiziert werden. Diese Elementespezifikationen werden mit `<elementSpec>` und `<attList>` (für die Attribute der Elemente) in der ODD umgesetzt. Ein Beispiel für eine Elementenspezifikation des Elementes `<extent>`[160] aus dem Module `biblPart` zeigt Abbildung 7.9.

[159] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-moduleRef.html` (besucht am 17.10.2016).

[160] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-extent.html` (besucht am 18.10.2016).

```
<TEI>
 <elementSpec ident="extent" module="header" mode="change">
       <classes mode="replace">
         <memberOf key="model.biblPart"/>
         <memberOf key="att.global"/>
         <memberOf key="att.typed" mode="add"/>
       </classes>
       <attList>
         <attDef ident="type" usage="" mode="change">
           <gloss>Indicates which unit of quantity you use.</gloss>
           <valList type="closed" mode="add">
             <valItem ident="Tokens">
               <gloss>The size of the corpus is given in token.</gloss>
             </valItem>
             <valItem ident="Words" mode="add">
               <gloss>The size of the corpus is given in words.</gloss>
             </valItem>
           </valList>
         </attDef>
         <attDef ident="ana" mode="delete"/>
         <attDef ident="change" mode="delete"/>
         <attDef ident="copyOf" mode="delete"/>
<!-- ... -->
         <attDef ident="xml:id" mode="delete"/>
         <attDef ident="xml:lang" mode="delete"/>
         <attDef ident="xml:space" mode="delete"/>
       </attList>
     </elementSpec>

</TEI>
```

Abbildung 7.9: *Beispiel für eine Elementspezifikation von* `<extent>` *für die Objekte der Klasse* ***Document****. Viele Attribute werden gelöscht, das* `@type` *wird verändert.*

Das Element `<extent>` ist Teil des `biblPart`-Modul der TEI[161] – `memberOf/@key`[162] Weiterhin erhält dieses Element seine Attribute aus der Attributgruppe *att.global*, das in dem `tei`-Modul enthalten ist. Die Attributgruppe *att.typed*[163] wurde diesem Element hinzugefügt, um eine Klassifizierung von `<extent>` zu ermöglichen. Diese Klassifizierung unterscheidet hier in einer geschlossen Attributwertliste, nach welchen Einheiten die Größe eines Objektes der Klasse **Document** angegeben wird – *Wörter* oder *Token*. Mit der Einheit *Token* kann die Anzahl der Token einer Annotationsebene mit eigenständiger Segmentierung angegeben werden. Die Einheit

[161] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-model.biblPart.html` (besucht am 16.10.2016).

[162] Diese Angabe ist hier explizit in der Spezifikation angegeben. Dies ist bei anderen Elementen, die keine Änderungen auf dieser Ebene besitzen, nicht der Fall.

[163] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-att.typed.html` (besucht am 18.10.2016).

Wörter ist eine konventionelle Angabe, die sich nicht auf die kleinsten technischen Einheiten im Korpus (Token) stützt. Sie ist als Alternative angegeben, wenn ein Korpus keine Tokenisierung besitzt. Diese Größen sind für die Wiederverwendung ganzer Korpora oder Teilen von Korpora relevant, wenn beispielsweise automatische Verfahren nur ab einer bestimmten Korpusgröße gut funktionieren.

Die Objekte der Klasse **Document** werden in einem `teiHeader` realisiert, der die strukturellen, administrativen, deskriptiven und technischen Metadaten einer Annotation eines Korpus enthält. Abbildung 7.10 zeigt die grobe Struktur des `teiHeader` ohne die Anzeige der tieferen XML-Struktur oder einzelner Werte.

Ein typischer TEI-konformer `teiHeader` besteht aus den Elementen `<fileDesc>` mit `<titleStmt>`, `<publicationStmt>`, `<sourceDesc>` sowie aus `<encodingDesc>` und `<revisionDesc>`. Die `<fileDesc>` „enthält die detaillierte bibliografische Beschreibung einer elektronischen Datei“[164], die hier auf ein Objekt der Klasse **Document** bezogen wird, das typischerweise bibliographische Beschreibungen erhalten kann. `<encodingDesc>` „dokumentiert das Verhältnis zwischen dem elektronischen Text und seiner Quelle oder den Quellen, von der oder von denen er abstammt“[165]. Zum Zweck der Darstellung des MKM in TEI wird `<encodingDesc>` nun so uminterpretiert, dass sie das Verhältnis zwischen Text und Umsetzung in Abhängigkeit der Objekte der Klasse **Annotation** beschreibt, da die Objekte der Klasse **Document** nicht unabhängig von den Objekten der Klasse **Annotation** existieren. Die `<revisionDesc>` „enthält alle Revisionsschritte, die an der Datei vorgenommen wurden.“[166]. Auf die Klasse **Document** uminterpretiert, enthält `<revisionDesc>` alle Revisionsschritte, die bezogen auf die Annotationen, die einem Text zugeordnet werden, durchgeführt worden sind.

Jedes Objekt der Klasse **Document** wird in einer TEI-XML-Dateien realisiert und die Metadaten zum Namen, den Herausgebern und den Autoren sowie der Veröffentlichungs- und Revisionsgeschichte werden umgesetzt (vgl. Abbildung 7.1).

[164] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-fileDesc.html` (besucht am 19.10.2016).
[165] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-encodingDesc.html` (besucht am 19.10.2016).
[166] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-revisionDesc.html` (besucht am 19.10.2016).

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader type="DocumentHeader">
     <fileDesc>
       <titleStmt>
         <title>...</title>
         <editor>...</editor>
         <author>...</author>
       </titleStmt>
       <extent>...</extent>
       <publicationStmt>
         <idno>...</idno>
         <pubPlace>...</pubPlace>
         <publisher>...</publisher>
         <date>...</date>
         <biblScope>...</biblScope>
       </publicationStmt>
       <seriesStmt>...</seriesStmt>
       <sourceDesc>...</sourceDesc>
     </fileDesc>
     <profileDesc> ... </profileDesc>
     <encodingDesc>
       <schemaSpec ident="AnnotationKey">
         <elementSpec ident="...">
           <valList>
             <valItem ident="..." corresp="..."/>
             <valItem ident="..." corresp="..."/>
           </valList>
         </elementSpec>
       </schemaSpec>
       <schemaSpec ident="AnnotationKey">
         <elementSpec ident="...">
           <valList>
             <valItem ident="..." corresp="..."/>
             <valItem ident="..." corresp="..."/>
             <valItem ident="..." corresp="..."/>
           </valList>
         </elementSpec>
       </schemaSpec>
     </encodingDesc>
     <revisionDesc> ... </revisionDesc>
   </teiHeader>
   <text/>
</TEI>
```

Abbildung 7.10: *Beispiel für Realisierung der TEI-Spezifikation für ein Objekt der Klasse* ***Document****.*

Neben den deskriptiven Angaben zu Titel, Autoren und Herausgebern und zur Publikationsgeschichte inklusiver aller möglichen bibliographischen Hintergrundinformationen zur Hand- und Niederschriften eines historischen Texts – `<fileDesc>` – werden pro Objekt der Klasse **Document** mit Hilfe der Referenzen angegeben, welche Annotationen enthalten sind. Hierfür wird wieder ähnlich wie bei Objekten

der Klasse **Annotation** die `<encodingDesc>` genutzt. Dieses Element enthält die technischen Metadaten – `<schemaSpec>` – zu den Annotationen, die sich ein Set an **Document**-Metadaten teilen.

Das Element `<schemaSpec>` hat die Funktion, ein TEI-konformes Schema z. B. in der ODD zu erzeugen.[167] Zum Zweck der Darstellung des MKM in TEI wird das Element `<schemaSpec>` nun so uminterpretiert, dass es ein Schema für Annotationen erzeugt bzw. abbildet. Annotationsrichtlinien sind ebenfalls Schemata, nach denen ein Korpus annotiert wurde oder beispielsweise im Fall eines Szenario 4 (Größenanreichung) neues sprachliches Material nach demselben Schema annotiert werden soll.

Eine für die bibliographischen Angaben zu den historischen Texten wesentliches Elementegruppe ist `<msDesc>`[168], die innerhalb von `<sourceDesc>` realisiert wird. Hier finden sich Metadaten zur den originalen Veröffentlichungszeitraum, -ort und Autoren eines historischen Texts, der beispielsweise in einer modernen Edition publiziert worden ist.

Wie bei der **Annotation**-Spezifikation ist der Aufbau der TEI-XML-Datei im Bereich des `teiHeader` mit dieser hier vorgestellten ODD-Spezifikation definiert (vgl. Abschnitt 7.1.1).

7.1.3 Spezifikation für die Klasse Corpus

Für ein Objekt der Klasse **Corpus** werden die TEI-Module `textstructure`, `core`, `header` und `namesdates` genutzt.[169] Die Anpassung dieser Module basiert ebenfalls hauptsächlich auf zwei Prinzipien: die Einschränkung der nutzbaren Elemente und deren Attribute sowie der Hinzufügung einiger weniger Elemente. Die Modulspezifikationen geben an, welche Module und welche Elemente oder Elementgruppen in welchen Modulen in der **Corpus**-TEI-Spezifikation genutzt werden, wie der folgende Auszug aus der ODD in Abbildung 7.11 zeigt.

[167] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-schemaSpec.html` (besucht am 15.01.2017).

[168] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-msDesc.html` (besucht am 18.10.2016).

[169] Dass Korpora auch Subkorpora beinhalten können, konnte bislang nicht in der TEI-MKM-Realisierung integriert werden.

```
<TEI>
<schemaSpec ident="teiODD_LAUDATIOCorpus_S7.1">
    <moduleRef key="core" include="author editor date list item label p ref title"/>
    <moduleRef key="tei"/>
    <moduleRef key="header"
        include="appInfo application authority availability change editorialDecl
        encodingDesc extent fileDesc idno langUsage language namespace normalization
        profileDesc projectDesc publicationStmt revisionDesc segmentation
        sourceDesc tagUsage tagsDecl teiHeader titleStmt"/>
    <moduleRef key="textstructure" include="TEI text"/>
    <moduleRef key="namesdates" include="affiliation persName forename surname orgName"/>
    <!-- ... -->
</schemaSpec>
</TEI>
```

Abbildung 7.11: *Ausschnitt aus der TEI-Spezifikation für die Objekte der Klasse **Corpus**.. Hier werden alle verwendeten Module der TEI sowie alle zusätzlich hinzugefügten Elemente aufgelistet.*

Die eingesetzten Module werden in der ODD jeweils mit `<moduleRef>`[170] aufgelistet, dessen Attribut *include* eine Liste aller Elemente angibt, die aus anderen Modulen kopiert und in das zu definierende Schema integriert werden. Keines der verwendeten Module wurde in seiner Struktur grundlegend verändert, es werden nur benötigte Elemente hinzugefügt. Die Elemente wiederum werden hinsichtlich ihrer Attribute verändert. z. B. werden häufig nicht benötigte Attribute entfernt oder einige Attribute erhalten offenen oder geschlossene Wertelisten. So sind beispielsweise alle nicht benötigten Attribute in der Spezifikation entfernt worden und andere Attribute des Elements mit festen Werten ausgestattet worden, wie z. B. bei dem Element `<author>` (Abbildung 7.12).

[170] `http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-moduleRef.html` (besucht am 17.10.2016).

```
<TEI>
        <elementSpec ident="author" module="core"
mode="change">
          <attList>
            <attDef ident="role" usage="rec" mode="change">
              <valList type="closed" mode="add">
                <valItem ident="Annotator"/>
                <valItem ident="Infrastructure"/>
                <valItem ident="Transcription"/>
              </valList>
            </attDef>
            <attDef ident="n" usage="rec" mode="change">
              <valList type="open" mode="change"/>
            </attDef>
            <attDef ident="ana" mode="delete"/>
            <attDef ident="change" mode="delete"/>
            <attDef ident="copyOf" mode="delete"/>
            <attDef ident="exclude" mode="delete"/>
            <attDef ident="corresp" mode="delete"/>
            <attDef ident="key" mode="delete"/>
          <!-- ... -->
            <attDef ident="style" mode="delete"/>
            <attDef ident="synch" mode="delete"/>
            <attDef ident="xml:base" mode="delete"/>
            <attDef ident="xml:id" mode="delete"/>
            <attDef ident="xml:lang" mode="delete"/>
            <attDef ident="xml:space" mode="delete"/>
          </attList>
        </elementSpec>
</TEI>
```

Abbildung 7.12: *Beispiel für eine Elementspezifikation des Elementes* `<author>` *für die Objekte der Klasse* ***Corpus****. So wurden beispielsweise viele Attribute gelöscht, das Attribut* `@role` *mit einer geschlossenen Werteliste ausgestattet.*

Hier wird neben der Löschung nicht gebrauchter Attribute das Attribut `@role` mit einer geschlossenen Attributwertliste – *Annotator*, *Instrastructure* und *Transcription* – versehen, was auch in allen anderen TEI-Spezifikationen umgesetzt ist. Mit dieser Typenspezifikation können nun Autoren innerhalb der TEI-Welt als Annotatoren und Transkriptoren sowie als Verantwortliche für die Infrastruktur des Korpus ausgewiesen werden. Diese Metadaten, neben den Metadaten zu den Herausgebern des Korpus, sind wichtig, um Ansprechpersonen für die verschiedenen Bereiche zu identifizieren.

Die Objekte der Klasse **Corpus** werden in einem `teiHeader` realisiert, der die strukturellen, administrativen, deskriptiven und technischen Metadaten zum Korpus enthält. Abbildung 7.13 zeigt die grobe Struktur des `teiHeader` ohne die Anzeige

der tieferen XML-Struktur oder einzelner Werte.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader type="CorpusHeader">
    <fileDesc>
      <titleStmt>
        <title>...</title>
        <editor>...</editor>
        <author>...</author>
      </titleStmt>
      <extent>...</extent>
      <publicationStmt>
        <authority>...</authority>
        <idno>...</idno>
        <availability>...</availability>
        <date>...</date>
      </publicationStmt>
      <sourceDesc>...</sourceDesc>
    </fileDesc>
    <profileDesc> ... </profileDesc>
    <encodingDesc> ... </encodingDesc>
    <revisionDesc> ... </revisionDesc>
  </teiHeader>
  <text/>
</TEI>
```

Abbildung 7.13: *Beispiel für Realisierung der TEI-Spezifikation für ein Objekt der Klasse* ***Corpus****.*

Ein typischer TEI-konformer `teiHeader` besteht aus den Elementen `<fileDesc>` mit `<titleStmt>`, `<publicationStmt>`, `<sourceDesc>` sowie aus `<encodingDesc>` und `<revisionDesc>`. Die `<fileDesc>` „enthält die detaillierte bibliografische Beschreibung einer elektronischen Datei“[171], die hier auf ein Objekt der Klasse **Corpus** bezogen wird, das ebenfalls als eine Art der Publikation bibliographische Beschreibungen erhalten kann. Die `<encodingDesc>` „dokumentiert das Verhältnis zwischen dem elektronischen Text und seiner Quelle oder den Quellen, von der oder von denen er abstammt“[172]. Zum Zweck der Darstellung des MKM in TEI wird das Element `<encodingDesc>` nun so uminterpretiert, dass es das Verhältnis zwischen dem Korpus und seiner Umsetzung in Abhängig der Objekte der Klassen **Document** und **Annotation** beschreibt, da die Objekte der Klasse **Corpus** nicht unabhängig von den Objekten der Klasse **Document** und diese wieder nicht unabhängig von Objekten

[171] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-fileDesc.html` (besucht am 19.10.2016).

[172] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-encodingDesc.html` (besucht am 19.10.2016).

der Klasse **Annotation** existieren. So werden die Objekte der Klassen **Annotation** und **AnnotationValue** in `encodingDesc` aufgeführt. Die `<revisionDesc>` „enthält alle Revisionsschritte, die an der Datei vorgenommen wurden. “[173]. Auf die Klasse der **Corpus** uminterpretiert, enthält `<revisionDesc>` alle Revisionsschritte, die in Bezug auf die Annotationen und damit auch auf die einzelnen Dokumente durchgeführt worden sind.

Jedes Objekt der Klasse **Corpus** wird in einer TEI-konformen Datei realisiert und die Metadaten zum Namen, den Herausgebern sowie der Veröffentlichungs- und Revisionsgeschichte werden umgesetzt (vgl. Abbildung 7.1).

7.1.4 Verbindung der Spezifikationsdokumente

Die Realisierung des MKM wird durch drei TEI-Spezifikationen durchgeführt. Eine wesentliche Motivation dafür ist, dass jeweils diese drei Klassen für das MKM zentral sind, diese jeweils eigene Objekte beschreiben. Die Metadaten der jeweiligen Objekte können mit der `teiHeader`-Struktur realisiert werden.

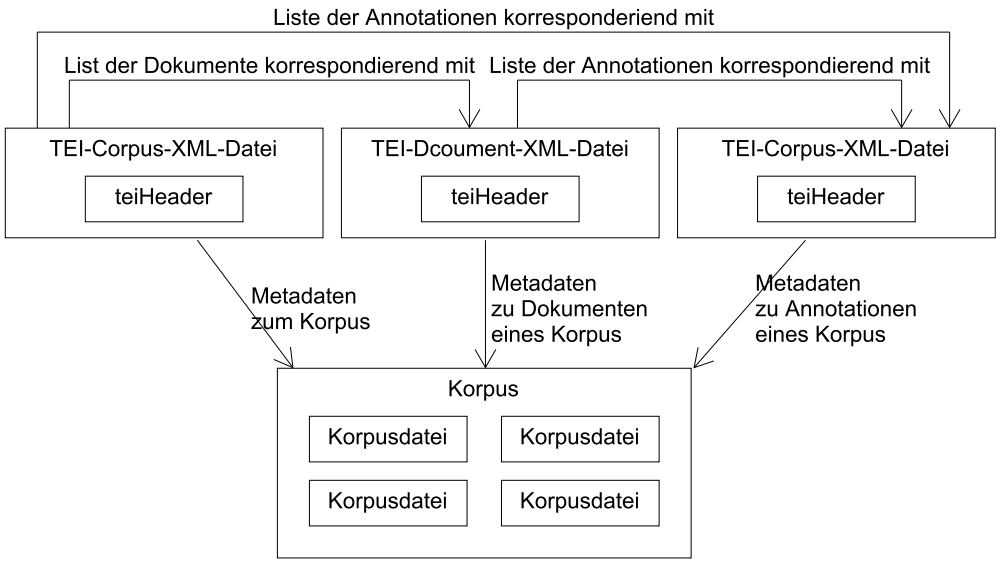

Abbildung 7.14: *Referenzierungen zwischen den TEI-Spezifikationen.*

Die Realisierung richtet sich damit nicht nach der vorhandenen Dateien des Korpus. Unabhängig davon, wie viele Dateien und in wie vielen Formaten ein Korpus

[173] `http://www.tei-c.org/release/doc/tei-p5-doc/de/html/ref-revisionDesc.html` (besucht am 19.10.2016).

vorhanden ist, die durch das MKM beschriebenen Klassen lassen sich davon unabhängig identifizieren. Die typische `teiHeader`-Struktur muss für eine Realisierung wenig uminterpretiert und angepasst werden, wieAbschnitt 7.1.1, Abschnitt 7.1.2 und Abschnitt 7.1.3, gezeigt haben.

Um die Verbindungen zwischen den einzelnen Klassen, die im MKM abgebildet sind, in den Realsierungen abzubilden, sind in allen TEI-Spezifikationen Referenzen über das Attribut *xml:ID*[174] eingebaut. Mit diesen Referenzen können Verbindungen zwischen den einzelnen `teiHeader` aufgebaut und ausgelesen werden, was Abbildung 7.14 illustriert. Die TEI-MKM-Spezifikation für Korpora enthält referenzierte Listen aller im Korpus enthaltenen Objekte der Klassen **Document** und **Annotation**. Die TEI-MKM-Spezifikation für Dokumente enthält eine referenzierte Liste aller Objekte der Klasse **Annotation**.

7.2 Anwendung für die TEI-Spezifikationen

Die Realisierung des MKM mit dem TEI-Metadatenmodell (nachfolgend TEI-MKM-Metadaten) kann in Anwendungen, die Korpora und Korpusdokumentationen verwaltet, eingesetzt werden. Mit MKM-TEI-Metadaten können jeweils Korpora mit den für alle Wiederverwendungsszenarien notwendigen Metadaten beschrieben werden. Damit sind bereits Handlung 1 (Deskription) und Handlung 4 (Authentifizierung) aus Sicht der Korpuserstellerin oder des Korpuserstellers (Akteurin/Akteur 1) ermöglicht (vgl. auf Metadaten basierende Handlungen Abschnitt 4.5). Für alle weitere Handlungen basierend auf Metadaten – Management, Retrieval, Interoperabilität und auch Deskription – werden Anwendungen benötigt, die die Metadaten vieler Korpora auslesen und auch für andere Forscherinnen und Forscher in einem gemeinsamen Zugang nutzbar machen können. Daher werden die drei TEI-MKM-Spezifikationen in einer Anwendungen, die die Wiederverwendung von Korpora ermöglichen, implementiert, dem LAUDATIO-Repository.[175]

LAUDATIO[176] ist ein Open Access Repositorium zur persistenten Speicherung von historischen Korpora (Krause et al. 2015). Bislang enthält LAUDATIO Korpora aus verschiedenen linguistischen Fachbereichen Korpora wie das Referenzkorpus Altdeutsch, das RIDGES-Korpus, das GerManC, das Mercurius-Korpus und das Mannheimer Korpus Historischer Zeitungen (vgl. Kapitel 2). Histori-

[174] `https://www.w3.org/TR/xml-id/` (besucht am 15.01.2017).

[175] Die Entwicklung der Software und der Arbeitsumgebung des Repositoriums sind nicht Teil der vorliegenden Arbeit.

[176] `http://www.laudatio-repository.org/` (besucht am 26.01.2017).

sche Korpora werden aus bereits abgeschlossenen wie laufenden Projekten in das Repositorium aufgenommen. Darüber hinaus ist das Repositorium offen für jede Art von historischem Korpus aus weiteren, andere Fachbereichen. LAUDATIO ist zusätzlich in einen integralen Ansatz zur Unterstützung des Datenmanagment der Humboldt-Universität zu Berlin eingebettet (Dreyer und Vollmer 2016).

Korpuserstellerinnen und -ersteller (Akteurin/Akteur 1) können mit Hilfe des Repositorium und der TEI-MKM-Metadaten ihre historische Korpora archivieren, referenzieren und dokumentieren. Jedes Korpus wird einheitlich, umfangreich, erstellerunabhängig mit den TEI-MKM-Metadaten dokumentiert und erhält eine Handle-PID[177]. Damit wird es den Korpuserstellerinnen und -erstellern möglich, ihre Datenmanagementpläne (DiPersio et al. 2016) insbesondere die langfristige Speicherung und Dokumentation ihrer Forschungsdaten über LAUDATIO umzusetzen. Nutzerinnen und Nutzer des Repositoriums (z. B. Akteurin/Akteur 3) können über eine Metadatenfacetten- und Freitextsuche auf Basis der jeweiligen MKM-TEI-Metadaten nach Korpora suchen.

Die TEI-MKM-Metadaten eines jeden Korpus werden in Form von TEI-XML-Dateien zusätzlich zu den eigentlichen Korpusdateien importiert und danach eingelesen. Da das Repositorium bislang nur eine Metadatendatei einliest, müssen die TEI-XML-Dateien zusammengefügt werden. Um dies für die Nutzerinnen und Nutzer leicht zu gestalten, wurde das Mergingtool *teitool*[178] entwickelt, dass die einzelnen TEI-XML-Dateien in eine – nicht mehr TEI-konforme – XML-Datei zusammenfügt.[179]

Die TEI-MKM-Metadaten werden dann für die verschiedenen Funktionen des Repositoriums auf folgende Weise eingesetzt: Es wird definiert, welche Elemente und Attribute mit welchen Werten wie in der Oberfläche des Repositoriums angezeigt werden sollen. Die MKM-TEI-Metadaten eines jeden Korpus werden im Repositorium einheitlich und tief strukturiert angezeigt. Die Nutzerinnen und Nutzer können sich über vier verschiedenen Metadatenrubriken Informationen über das jeweilige Korpus, die enthaltenen Dokumente und Annotationen sowie über die Aufbereitung einholen. Die Beziehungen der einzelnen Objekte der Klassen im MKM wird bereits in der TEI-Spezifikation zu einem großen Teil übernommen und kann so auch im Repositorium als zugrunde liegende Struktur dienen (Abbildung 7.15).

[177] `http://handle.gwdg.de:8080/pidservice/` (besucht am 19.10.2016).
[178] `https://github.com/thomaskrause/laudatioteitool` (besucht am 15.01.2017).
[179] Für die Dokumentation siehe `https://github.com/thomaskrause/laudatioteitool` (besucht am 20.10.2016).

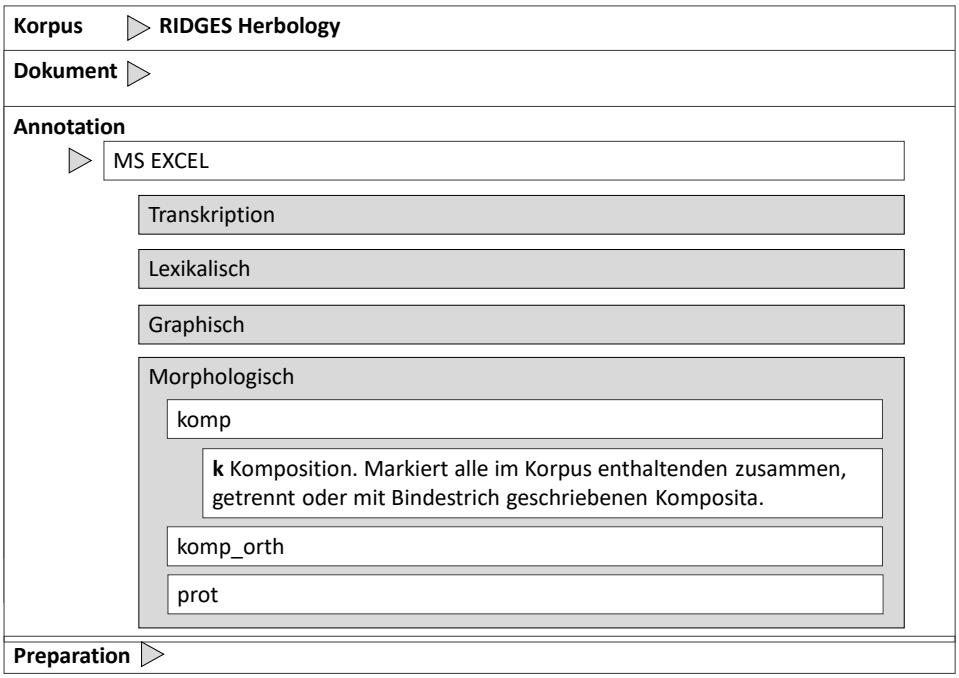

Abbildung 7.15: *Skizze einer Korpusdokumentation für ein Korpus in LAUDATIO. Der Anzeige der Korpusmetadaten liegt die Struktur zugrunde, die im MKM über die Beziehungen zwischen den Klassen modelliert ist. Abgebildet sind Metadaten zur Annotation komp in RIDGES.*

Abbildung 7.15[180] zeigt die Metadaten der Annotationsebene *komp* in RIDGES, die bereits in Abschnitt 7.1.1 kurz vorgestellt wurde. In der Korpusdokumentation werden vier große Informationsbereiche (Rubriken) für jedes Korpus angezeigt: Korpus, Dokument, Annotation, Preparation. Die Rubrik *Annotation*, die den Objekten der Klassen **Annotation** und **AnnotationValue** entspricht, enthält beispielsweise für jedes Format eine Auflistung der Menge aller im Korpus vorkommenden Annotationen mit ihren jeweiligen Werten und Beschreibungen. Diese sind in groben der Orientierung dienenden Gruppen eingeteilt (z. B. lexikalisch, morphologisch, graphisch). Für jede Annotation werden in der Rubrik *Preparation* alle Verarbeitungsschritte einzeln aufgelistet. Zusätzlich wird für jedes Korpus aus den TEI-MKM-Metadaten eine Zitiervorlage erstellt.

Damit Nutzerinnen und Nutzer in der vorhandenen Menge an Korpora im Repositorium suchen können, werden die TEI-MKM-Metadaten in eine Freitextsuche integriert. Weiterhin werden zentrale Metadaten in eine Facettensuche indexiert

[180] Die Skizze ist für diese Arbeit auf Deutsch gestaltet. Die Umsetzung der Anzeige im Repositorium ist hingegen auf Englisch.

(Abbildung 7.16). Einzelne Werte der TEI-MKM-Metadaten wie beispielsweise die Korpusname, die Korpusgröße, die enthaltenen Annotationen und Formate sind wesentliche Informationen, nach denen eine Menge an Korpora gefiltert werden kann. Für eine Anreicherung eines Korpus mit Annotationen (Szenario 3) ist beispielsweise die Information wesentlich, in welchen Formaten Korpora vorliegen, da die neu zu annotierenden Annotationskonzepte (und ggf. ihre Kategorien) Anforderungen an die Formate stellen. Oder die Zusammenführung von unterschiedlichen Annotationskonzepten ist von bestimmten Formaten (oder Tools) abhängig. Für eine Analyse eines Korpus ist diese Angabe ebenfalls wesentlich, da nicht jedes Format von einem Analyse oder Such- und Visualisierungstool eingelesen werden kann. In beiden Fällen können die Nutzerinnen und Nutzer nicht geeignete Kandidaten aus der Menge an Korpora mit dieser Metadatenfacette herausfiltern.

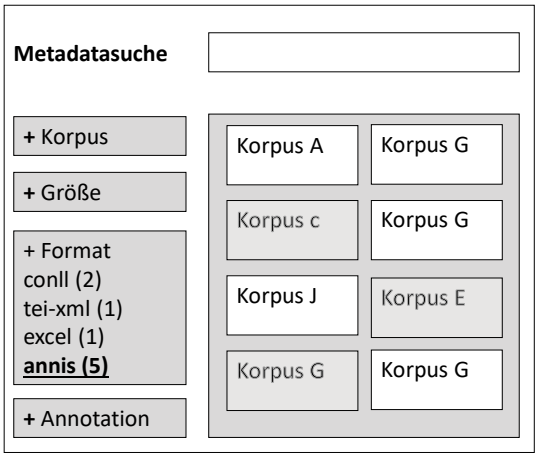

Abbildung 7.16: *Skizze der Metadatenfacettensuche in LAUDATIO. Jede Annotation besitzt die Angabe, in welchem Format sie vorliegt. Diese Angaben werden in einer Facette Format in Bezug auf Korpora zusammengefasst. Die Zahl hinter dem konkreten Wert der Metadatenfacette Format zeigt dann an, wie viele Korpora im Repositorium enthalten sind, die in einem bestimmten Format vorliegen.*

Abbildung 7.16 illustriert die Metadatenfacettensuche. Jede Annotation in allen Korpora hat in den TEI-MKM-Metadaten die Angabe erhalten, in welchem Format sie vorliegt. Diese Metadaten werden in der Facette *Format* über die Korpora, die die entsprechenden Annotationen enthalten, zusammengefasst. Die Auswahl von einem oder mehreren Formaten in dieser Facette schränkt dann die gesamte Treffermenge der Korpora ein. Die Facette drückt mit ihren Werten und den angegebenen Vorkom-

men aus, dass es *fünf Korpora* gibt, die in dem *annis*-Format vorliegen. Diese sind in Abbildung 7.16 weiß hinterlegt, Korpora, auf die dieses Merkmal nicht zutrifft, grau. Dieser Filtermechanismus hat den Vorteil, dass für Nutzerinnen und Nutzer direkt alle möglichen Werte einer Facette angezeigt werden und sie aus dem vorhandenen Angebot wählen können. Damit wird keine tiefer gehende Kenntnis über die Metadaten der im Repositorium vorhandene Korpora und damit über das zugrunde liegende Modell vorausgesetzt, die Struktur des Modells aber in der Anwendung ausgenutzt.

So können die Metadaten, die mit dem MKM modelliert und über die TEI-Modelle realisiert werden, in verschiedenen Funktionen und Mechanismen in einem Repositorium für historische Korpora genutzt werden. In wie weit das MKM und seine Realisierung die in Abschnitt 4.7 beschriebenen Qualitätsprinzipien adressiert oder erfüllen kann, soll kurz im nachfolgenden Abschnitt diskutiert werden.

7.3 Qualitätsprinzipien

Der Qualitätsbegriff selbst ist bereits viel diskutiert und kann auf unterschiedliche Weisen überprüft werden. Die Qualität von Metadaten kann dabei in Prinzip auf drei Perspektiven aufbauen, die der Korpuserstellerinnen und -ersteller, die der Korpusnutzerinnen und -nutzer und allen Personen oder Institutionen, die die Korpora verwalten und zur Verfügung stellen (Abschnitt 4.7). Dieser Perspektivenwechsel wurde bereits am Anfang der Entwicklung des MKM berücksichtigt, in dem vorher die verschiedenen Akteurinnen und Akteuren definiert und mit den Handlungen auf Basis von Metadaten sowie den Wiederverwendungsszenarien verknüpft wurden (vgl. Abschnitt 6.3 und Abschnitt 7.2).

In dieser Arbeit wird eine qualitative Einschätzung der Metadaten nach dem MKM auf Grundlage der sechs Qualitätsprinzipien nach NISO (2007) erarbeitet. Hervorgehoben werden soll an dieser Stelle, dass hier die Qualität der Metadaten nur in Bezug auf den vorher definierten Zweck und den zu beschreibenden Korpustyp erörtert wird. Weiterhin werden die Qualitätsmerkmale in Bezug auf das MKM und seine Realisierung, nicht aber auf die konkreten Korpusmetadaten, die Nutzerinnen und Nutzer vergeben, und auf die beschriebenen Forschungsdaten thematisiert, da sich die Qualität der Korpusdaten je nach Wiederverwendungsszenario und Forschungsfrage unterschiedlich bewerten lässt. Darüber hinaus werden erste Ansätze zur Unterstützung in der Anwendung sowie Erweiterung der Funktionen des MKM durch andere Systeme vorgestellt.

Das MKM ist an der UML angelehnt entwickelt worden und folgt damit einer etablierten Methode aus der Informatik. Die Realisierung des MKM durch die TEI folgt einem überfachlich akzeptierten und genutzten Defacto-Standard (adressiert das Metadata Principle 1). Weiterhin ist die hier vorgestellte Realisierung ein originäres Subset der TEI, dass formal durch die ODD-Spezifikation modelliert ist, und folgt damit den Anpassungs- und Interoperabilitätsstandards der TEI. Damit sind auch die Realisierung des MKM in die Modellwelt der TEI eingebettet. Das MKM ist nicht abhängig von einer Art Realisierung und kann daher weitere andere Realisierungen in beispielsweise CMDI erhalten (adressiert Metadata Principle 2+6). Das MKM gibt objekt- und zweckbezogen klare Strukturen und Klassen vor, die es ermöglichen, die jeweilig beschriebenen Objekte zu archivieren, zu verwalten und administrative Metadaten zu Verantwortlichen und Lizenzen anzugeben. Die TEI-MKM-Metadaten werden auch in einem Open Access Repositorium, das historische Korpora langfristig und nachhaltig zur Verfügung stellt, eingesetzt (adressiert das Metadata Principle 5). Die Metadaten dokumentieren weiterhin ein Korpus so, dass es unabhängig von den Erstellerinnen und Erstellern erschlossen und weiterverarbeitet werden kann. So leisten die Metadaten einen Beitrag zur langfristigen Speicherung von Korpora. Weiterhin werden auch klare Angaben zu den Nutzungsbedingungen gemacht (adressiert das Metadata Principle 4). Das MKM richtet sich für die Beschreibung der Korpora zu einem Teil nach den Inhaltsstandards der FRBR und der TEI. Zu einem anderen Teil abstrahiert es aus den vorhandenen Datenstrukturen der Korpora wie die Tokenisierung und Annotationskonzepte, die in Formaten realisiert sind, und fachübergreifend genutzt werden (adressiert das Metadata Principle 3).

Das MKM und seine Realisierung adressieren damit die National Information Standards Organization (NISO)-Metadatenprinzipien. Die TEI-MKM-Metadaten unterstützen die verschiedenen Akteurinnen und Akteure bei den verschiedenen Aufgaben des Datenmanagements, die wiederum eine Voraussetzung für ihre Wiederverwendung darstellt.

Ein wesentlicher Aspekt ist die Nutzbarkeit des Metamodells und seiner Realisierung für die Akteurinnen und Akteure selbst; gemeint ist die Erstellung der Metadaten (im Vergleich zur Nutzung der Metadaten in einem Repositorium, Abschnitt 7.2). Das MKM abstrahiert über Korpusarchitekturen von historischen Korpora und begreift ein Korpus als Menge seiner Dokumente und die Dokumente wiederum als Menge ihrer Annotationen. Ein Korpus wird weiterhin als Ergebnis (Produkt) aus mehreren Bearbeitungsschritten abgebildet (Forschungsdatenzyklus). Damit kann

die Erstellung der Metadaten an die Erstellungsprozesse des Korpus selbst gekoppelt werden, um den Nutzerinnen und Nutzern die Erstellung der Metadaten zu erleichtern. Mit der indirekten Einbindung der TEI-Realisierung in ein Konverterframework kann dies beispielsweise ermöglicht werden: Das Konverterframework PEPPER besitzt mit dem INFOMODULE die Möglichkeit, bestehende Annotationen mit den verwendeten Tags in einem Korpus pro Dokument auszulesen und in strukturierter Form aufzulisten (Voigt et al. 2016). Wenn ein Korpus in einem oder mehreren Formaten vorliegt, das mit einem PepperModule bereits unterstützt wird, dann erfolgt nach dem Einlesen des Formats in das Metamodell *Salt* ein Export der ausgelesenen Metadaten (in Abhängigkeit vom eingelesenen Format). Dieser strukturierte Output (des Exportvorgangs) wird durch ein Mapping-Prozess in die TEI-MKM-Spezifikationen überführt. Für das Schema Version 7 der TEI-Spezifikationen existiert ein Mapper, der den Output des InfoModules direkt in in TEI-konforme Dateien gewandelt.[181] Mit dieser Verknüpfung können bereits viele Metadaten der Objekte der Klasse **Annotation** und **AnnotationValue** sowie administrative und strukturelle Metadaten zu Objekten der Klassen **Document** und **Corpus** automatisch ausgelesen werden.

Neben der Wiederverwendung der historischen Korpora können die Metadaten selbst, wie sie in dieser Arbeit entwickelt sind, auf eine zweite Wiese verwendet werden, da sie eigenständig methodisches Wissen langfristig sichern und zur Verfügung stellen. Die Metadaten können dann als eine Art Bauplan für Korpora fungieren, die auch für die Erstellung von anderen Korpora genutzt werden können. Durch die tiefe Strukturierung der Modellierung ist es beispielsweise möglich, dass sich alle Akteurinnen und Akteure über einzelne oder mehrere Instanziierungen von Klassen oder Gruppen von Klassen zu unterschiedlichen Themen informieren können: Annotationskategorien, die bereits auf die Weise dokumentiert sind, oder Metadaten eines Bearbeitungsschrittes wie die Nutzung eines Annotationstools oder eines Konvertierungstools können beispielsweise auf eine Anwendung in anderen Korpora geprüft werden.

Dadurch, dass sich die Metadaten hier nicht primär nach fachspezifischen Informationen und Eigenschaften von Korpora richten, können also das Wissen und die Informationen, die sie enthalten, wiederverwendet werden. Damit haben Akteurinnen und Akteure Zugriff auf das strukturierte Wissen über verschiedene Korpora und deren Erstellungsmethoden, Annotationsverfahren und -architekturen. Eine Metarecherche über solche Korpusdokumentationen befähigt Akteurinnen und Akteure,

[181] Frei verfügbar unter `https://github.com/korpling/IM2HeaderMapper` (besucht am 31.10.2016).

nach Architekturen, Annotationskonzepten und Erstellungsmethoden unabhängig von der Forschungsfrage der jeweiligen Korpora zu suchen und sich darüber zu informieren, wie vorhandene Lösungen welche Methoden eingesetzt haben.

Um die Qualität der Metadaten in einem weiteren Bezug zu testen, können Usability Studies die Anwendung der Metadaten im Repositorium untersuchen (vgl. z. B. Kirschenbaum 2004). Solche Studien können beispielsweise überprüfen, wie nutzerorientiert die Metadaten in eine Metadatensuche und -anzeige umgesetzt sind. Einen ersten Ansatz zeigen Stiller et al. (2016) am Beispiel von LAUDATIO. Diese Art der Qualitätsprüfung der Metadaten ist eng verknüpft mit der Softwarearchitektur des Repositoriums sowie der dessen Front-End-Design. Ein weiterer Aspekt, der die Qualität der Metadaten im Hinblick auf deren Anwendung unterstützen kann, ist die Entwicklung eines Metadateneditors für die TEI-MKM-Metadaten, der die Nutzbarkeit in Bezug auf die Erstellung der konkreten Metadaten für ein Korpus fördern kann. Beide Aufgaben werden als Desiderat verstanden, können aber nicht im Rahmen dieser Arbeit bearbeitet werden, da die Softwareentwicklung, -optimierung und deren Anpassung nicht Teil der Arbeit ist.

8 Zusammenfassung der Ergebnisse

Die vorliegende Arbeit befasst sich mit den Anforderungen an eine Korpusdokumentation, die als eine Voraussetzung für die Wiederverwendung von Korpora verstanden und als ein Bestandteil der Veröffentlichung und Archivierung von Korpora verwendet werden kann. Ein geeignetes Anwendungsbeispiel stellen historische Textkorpora dar, da sie in vielen Fächern als empirische Grundlage für die Forschung genutzt und vielfältig wiederverwendet werden können. Historische Korpora zeichnen sich im Weiteren durch starke Unterschiede in ihrer Aufbereitung und Annotation von historischen Texten und ein komplexes Verhältnis zu der historischen Vorlage aus, wodurch sie besondere Anforderungen an eine Korpusdokumentation stellen. Was müssen also Forscherinnen und Forscher über ihr Korpus mit Hilfe von Metadaten dokumentieren, um dessen Erschließung und Wiederverwendung für andere Forscherinnen und Forscher ermöglichen zu können? Welche Funktionen übernehmen dabei die Metadaten?

Eigenschaften historischer Korpora und deren Wiederverwendungsszenarien In Kapitel 2 wird herausgearbeitet, welche gemeinsamen Eigenschaften historische Korpora besitzen, die relevant für eine Korpusdokumentation zum Zweck der Wiederverwendung sind. Dazu müssen mehrere historische Korpora mit unterschiedlichen historischen Vorlagen und verschiedenen Korpusarchitekturen betrachtet worden. Historische Korpora besitzen wie alle Forschungsdaten einen eigenen Forschungsdatenzyklus, der sie u. a. als Ergebnis (Produkt) aus verschiedenen Bearbeitungsschritten versteht. Historische Korpora besitzen weiterhin wie andere Korpustypen eine spezifische Architektur, die sich aus der jeweiligen Tokenisierung, verschiedenen Annotationskonzepten und -kategorisierungen ergibt und die in verschiedene Formate umgesetzt werden kann. Korpora werden als Produkt des Forschungsprozesses verstanden, was z. B. auch die Vielzahl an unterschiedlichen Annotationskategorisierungen für verwandte oder gleiche Konzepte wie Wortart widerspiegelt. Diese Kategorisierungen sind stark korpusspezifisch und kontextspezifisch, dass hier keine allgemeinen Eigenschaften für eine Korpusdokumentation, wie z. B. *Wortartenanno-*

tation abgeleitet werden können. Die Annotationskategorien können in unterschiedlichen Annotationskonzepten wie Spannen oder Bäume abgebildet werden. Diese wiederum können über die Menge an betrachteten Korpora abstrahiert zusammengefasst und als ein Beschreibungsmerkmal identifiziert werden. Allgemein werden die jeweiligen Realisierungen dieser Annotationskonzepte in bestimmten Formaten mit bestimmten Bearbeitungsschritten und Tools erstellt. Formate und Tools besitzen vom Korpus unabhängige (korpusexterne) Eigenschaften. Welche Annotationen mit welchem Bearbeitungsschritt und welchem Tool durchgeführt werden, sind hingegen relevante Informationen für die Wiederverwendung, wenn beispielsweise eine weitere Annotation im Korpus hinzugefügt werden soll.

Für historische Korpora spezifisch ist die Beziehung zwischen der historischen Vorlage (dem Text) und dem Digitalisat (dem Korpus). Vergleichbar mit den Annotationskategorien existieren ganz unterschiedliche Ansätze, einen historischen Text zu digitalisieren. Dies zeigt sich in Form von verschiedenen Transkriptions- und Normalisierungsrichtlinien. Diese Richtlinien können nicht oder nur schwer aufeinander abgebildet werden, da sie jeweils ganz unterschiedliche zum Teil sich widersprechende Entscheidungen beinhalten, die sich aus dem Forschungsprozess heraus motivieren. So greifen auch Transkriptionen unterschiedlich normalisierend ein und es gibt keine klare Grenze, ab wann von einer Normalisierung gesprochen werden kann. Diese Entscheidungen beruhen auch darauf, welche historische Vorlage jeweils gewählt wurde. Die eindeutige Definition der historischen Vorlage ist ebenfalls schwierig. Dasselbe historische Werk kann in verschiedenen Expressionen und Manifestationen vorliegen. Jede Expression eines Werks kann eigene Besonderheiten besitzen, die untereinander verglichen werden können und damit nicht als dasselbe interpretiert werden. Welche Manifestation einer Expression oder auch welches Exemplar als Vorlage genutzt wird, muss daher in eine Korpusdokumentation aufgenommen werden. In jedem Korpus können so ganz unterschiedliche Textkonzepte integriert werden. In welcher Beziehung also die jeweiligen Transkriptionen oder Normalisierungen zu den jeweiligen Manifestationen oder Expressionen sowie zum Werk stehen, kann und soll nicht allgemein abgeleitet werden. Daher werden hier alle Arten von Transkription und Normalisierungen einheitlich als Annotationen verstanden und müssen in eine Korpusdokumentation auch so aufgenommen werden.

Weiterhin werden die unterschiedlichen Nutzergruppen und deren möglichen Wiederverwendungsszenarien von historischen Korpora in Kapitel 3 aufgeführt, wodurch der Zweck der Korpusdokumentation klar herausgestellt wird. Neben den Korpuserstellerinnen und -erstellern können auch andere, dritte Forscherinnen und Forscher

oder eine Gruppe aus Korpuserstellerinnen und -erstellern und dritten Forscherinnen und Forschern ein Korpus wiederverwenden. Diese jeweiligen Akteurinnen und Akteure besitzen unterschiedliche Kenntnisstände, die jeweils von einer Korpusdokumentation berücksichtigt werden müssen. Mit dem RIDGES-Korpus wird ein detaillierter Einblick in einen Forschungsdatenzyklus eines historischen Korpus gegeben, aus dem Wiederverwendungsszenarien abgeleitet werden. Unter der Wiederverwendung von Korpora werden in dieser Arbeit mehrere Szenarien definiert, die getrennt oder in Kombination auftreten können. So können Korpora neu oder erneut analysiert, mit weiteren Annotationen erweitert, um bestimmte Annotationen reduziert oder in andere Formate konvertiert werden.

Metadaten und Metadatenstandards In Kapitel 4 wird vorgestellt, welche unterschiedlichen Klassifikationen und Funktionen für Metadaten existieren, mit deren Hilfe eine Korpusdokumentation erstellt werden kann. Metadaten werden funktional in deskriptive, strukturelle, administrative und technische Metadaten eingeteilt. Für die Beschreibung eines Korpus als Ergebnis (Produkt) eines Forschungsdatenzyklus müssen die verschiedenen Metadaten die Korpuseigenschaften aus einer produktorientierten Perspektive beschreiben. Damit werden nur Informationen relevant, die das gespeicherte, nicht-flüchtige Korpus näher bestimmen. Es werden nur Bearbeitungsschritte und verwendete Tools dokumentiert, die zu dem Ergebnis – veröffentlichtes Korpus – geführt haben und damit von anderen Forscherinnen und Forschern nachvollzogen werden müssen. Nicht zu berücksichtigen sind so beispielsweise revidierte Bearbeitungsschritte, die dennoch Teil des Forschungsdatenzyklus sind.

Um ein Korpus wiederverwenden zu können, muss das Korpus erschlossen werden. Damit stiften die Wiederverwendungsszenarien den eigentlichen Zweck der Korpusmetadaten, die einer Erschließung der Korpora erstellerunabhängig ermöglichen sollen. Es existieren verschiedene Akteurinnen und Akteure, die unterschiedliche Kenntnisstände in Bezug auf die Korpora besitzen und damit ebenfalls unterschiedliche Anforderungen an die Metadaten stellen. Auf Grundlage der Metadaten können Akteurinnen und Akteure verschiedene Handlungen durchführen. Zuerst müssen initiale Korpuserstellerinnen und -ersteller ihre Korpora beschreiben und referenzierbar machen, was auf Metadaten basierende Handlungen darstellen. Erst danach können andere Forscherinnen und Forscher auf Grundlage von Korpusmetadaten nach Korpora suchen und gefundene Korpora über deren Metadaten erschließen. Es muss also auf Grundlage der Metadaten (in einer Anwendung) möglich sein, nach den relevanten Eigenschaften wie bestimmten Annotationskategorien und deren Verarbeitung

zu suchen bzw. sich pro Korpus darüber konkret zu informieren.

Daraufhin werden in Kapitel 5 bestehende Metadatenstandards diskutiert, die für die Dokumentation von Korpora bereits eingesetzt werden oder eingesetzt werden können. Hierzu werden die für eine Korpusdokumentation relevanten Eigenschaften in Beschreibungskomponenten zusammen gefasst (*Quelle, Inhalt/Struktur, Veröffentlichung, Erstellung/Bearbeitung*). Es wird gezeigt, dass die bisherigen Ansätze diese relevanten Korpuseigenschaften mit den notwendigen Metadaten für die unterschiedlichen Nutzergruppen nicht oder nicht vollständig in einem gemeinsamen Beschreibungsmodell abbilden (können).

Der allgemeine DublinCore-Standard mit seinen 15 Kernelementen und dessen angepassten Varianten besitzt einen zu geringen Informationsumfang und keine Struktierungsmöglichkeit der Elemente als Attribut-Wert-Paare, so dass nicht alle Beschreibungskomponenten in vollem Umfang und in ihrer tieferen Strukturierung erfasst werden können.

Der Ansatz der ISLE Meta Data Initiative (IMDI) besitzt ein umfangreicheres Elementeset und die Möglichkeit, die Metadaten zu strukturieren. IMDI ist hingegen für einen anderen Korpustyp, nämlich Korpora der gesprochene Sprache, entwickelt worden und kann die Anforderungen, die historische Korpora an eine Korpusdokumentation stellen, daher nicht gut adressieren. Der Ansatz der Component MetaData Infrastructure ist für verschiedene Korpustypen entwickelt worden und besitzt im Gegensatz zu den anderen Metadatenstandards kein festes Elementeset, da er sich als Framework versteht, mit denen Nutzerinnen und Nutzer ihre eigenen Metadaten erstellen können. Die CMDI definiert Metadaten und deren Granularität so, dass nur externe Eigenschaften einer Ressource direkt mit Metadaten beschrieben werden, interne Eigenschaften wie Annotationskategorisierungen und -konzepte sollten separat beschrieben werden. Diese Metadaten sind jedoch wesentlich für eine Korpusdokumentation von historischen Korpora (besonders die Beschreibungskomponenten *Inhalt/Struktur* und *Quelle*). Eine Anpassung bereits bestehender nutzerspezifischer CMDI-Schemata (Profile), entspricht so einerseits nicht den CMDI-Vorgaben und ist anderseits schwer möglich, da CMDI kein eigenes einheitliches Metamodell oder eine eigene abstrakte Beschreibungsebene besitzt. Jede CMDI-Nutzerin oder jeder CMDI-Nutzer verwendet implizit oder explizit eigene Modelle, so dass auf dieser Ebene unter Umständen konkurrierende, ähnliche oder nicht aufeinander abbildbare Bedeutungen in Profilen verwendet werden.

Der Ansatz des Metadata Encoding and Transmission Standard (METS) stützt sich auf eine dokumentenorientierte Beschreibungsebene, die besonders die

Anforderungen der verschiedenen Textdefinitionen berücksichtigt. Damit kann die Beschreibungskomponente *Quelle* umfangreich umgesetzt werden, weitere Informationen zu den Annotationen können hingegen nicht abgebildet werden. Eine Anpassung des Standards ist nicht möglich. Die Text Encoding Initiative (TEI) integriert mit ihrem ebenfalls dokumentorientierten Ansatz die Beschreibungskomponente *Quelle* in ihren Guidelines. Die Metadaten der TEI beziehen sich vorwiegend auf ein Dokument und nicht so sehr auf ein Korpus. Innerhalb der TEI-Welt ist eine derartig umfangreiche Dokumentation durch Metadaten auch nicht notwendig, da TEI-konforme Dateien den TEI-Guidelines bereits folgen. Die TEI wird daher selbst wenig als Metadatenstandard genutzt, sie besitzt aber im Vergleich zu den anderen Ansätzen einen integrierten Modellierungmechanismus in Form des Spezifikationsdokumentes TEI-Spezifikation One Document does it all (ODD), der eine Anpassung der TEI als ein Metadatenmodell über eine einheitliche abstrakte Modellierungssprache ermöglicht.

Es fehlt also grundsätzlich ein Konzept dafür, wie das Geflecht aus Metadaten zu Quellen, digitalen Surrogaten und Annotationen in einer Korpusdokumentation organisiert werden kann. Ein solches Konzept wird in Form eines Metamodells für Korpusmetadaten in dieser Arbeit vorgestellt. Es soll die notwendige abstrakte Beschreibungsebene zur Verfügung stellen.

Das Metamodell für Korpusmetadaten als gemeinsames Beschreibungsmodell für historische Korpora Das MKM modelliert Metadaten von historischen textbasierten Korpora aus einer technisch-abstrakten, produktorientierten und damit überfachlichen Perspektive. Dafür werden deskriptive, strukturelle, administrative und technische Metadaten benötigt, auf deren Basis verschiedene Akteurinnen und Akteure verschiedene Handlungen durchführen können. Solche Handlungen können das Beschreiben von Korpora, das Authentifizieren von Korpora, das Suchen nach Korpora oder ganz allgemein das Korpusdatenmanagement sein. Diese Handlungen sind dann die Voraussetzung für eine Wiederverwendung von Korpora. Das MKM ist damit zweckgebunden und adressatenbezogen.

Grundlage für die Entwicklung des MKM sind Abstrahierungen von Korpuseigenschaften und Korpusentwicklungen, die aus der Betrachtung einer Menge an historischen Korpora gewonnen werden. Die Arbeit zeigt, welche korpuseigene Eigenschaften und korpusexterne Eigenschaften sich historische Korpora aus verschiedenen Fächern teilen und wie sie mit Metadaten beschrieben werden können. Dafür werden zentrale Beschreibungsklassen für das Korpus, seine Dokumente und An-

notationen modelliert. Externe Metadaten können ebenfalls als Klassen abgebildet werden, die z. B. beteiligte Personen oder Institutionen sowie Bearbeitungsschritte und verwendete Tools beschreiben. Diese Klassen sind über verschiedene Arten von Beziehungen miteinander verbunden und bilden damit eine komplexe Metadatenstruktur für Korpora.

Ein Korpus besteht nach dem MKM aus mehreren Dokumenten und ein Dokument aus mehreren Annotationen. Diese Objekte werden jeweils zu Klassen abstrahiert, die jeweils mit folgenden Eigenschaften und weiteren Beziehungen zu Klassen beschrieben werden können.

Für die Modellierung der Klasse **Annotation** sind nicht die Interpretationen und Konzepte, die die Annotationen darstellen können, maßgeblich, sondern die strukturellen Relationen zwischen Annotationen sowie ihre Erstellung selbst. Da auch jede Form der Transkription und Normalisierung als Annotation modelliert wird, werden die auf Textebenen bezogenen Metadaten nicht in dieser Klasse, sondern über eine weitere Klasse **Document** repräsentiert. Alle Annotation werden durch einen oder mehrere Bearbeitungsschritte in einem oder mehreren Formaten erzeugt oder weiter bearbeitet. Ein Bearbeitungsschritt findet in einem oder mehreren Formaten statt. Dies wird im MKM in den Klassen **Preparation** und **Format** abgebildet, wobei jedes Objekt der Klasse **Annotation** aus einem oder mehren Objekten der Klasse **Preparation** besteht, und jedes Objekt der Klasse **Preparation** aus einem oder mehreren Objekten der Klasse **Format**. Wesentliche Eigenschaften der Annotationen, die mit dieser Struktur abgebildet werden können, sind die Angaben, welche Annotationskategorien in welchen Annotationskonzepten und Formaten umgesetzt sind. Mit der Angabe, welche Annotation in welchem Format eine abhängige oder eine eigenständige Segmentierung besitzt, wird ein wesentliches Merkmal der Korpusarchitektur dokumentiert. So können beispielsweise Mehrebenenkorpora dadurch beschrieben werden, dass sie eine Annotation besitzen, auf deren Segmentierung sich eine Menge an weiteren Annotationen – unabhängig von ihrem Annotationskonzept – bezieht. Die Eigenschaft von Korpora, multiple Segmentierungen zu enthalten, wird dann dadurch beschrieben, dass ein Korpus mehr als eine Annotation mit einer unabhängigen Segmentierung besitzt. Damit werden (hauptsächlich) die Beschreibungskomponenten *Erstellung/Bearbeitung* und *Inhalt/Struktur* berücksichtigt.

Ein Dokument – Klasse **Document** – besteht aus einer oder mehreren Annotationen und es trägt die Eigenschaften der historischen Vorlage. Mit historischer Vorlage kann auf ein Werk referiert werden, das in einem oder mehreren Manifestationen publiziert wird (Objekt der Klasse **Publication**). Weiterhin können einem

Dokument eine oder mehrere weitere Quellen zugewiesen werden (Objekt der Klasse **Source**), die in einer Beziehung zu demselben Werk stehen. So können die Edition eines Werks und dessen Handschriften dokumentiert werden, die entweder beide als historische Vorlage für ein Korpus dienen oder wovon die Edition als Vorlage des Korpus dient, diese selbst aber eine bestimmte Handschrift ediert. Damit wird die Beschreibungskomponente *Quelle* im MKM berücksichtigt, die u. a. auf der Vier-Level-Beschreibung der FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS (FRBR) basiert. Auf diese Weise hat der Dokumentbegriff, wie ihn das MKM nutzt, keinen festen definitorischen Bezug zu einem Textbegriff, der eine theoretische, teilweise inhaltsbezogene und fachspezifische Interpretation besitzt.

Ein Korpus – Klasse **Corpus** – wiederum besteht aus einem oder mehreren Dokumenten. Die Metadaten von Objekten der Klasse **Corpus** sowie die mit ihnen assoziierten Objekten der Klassen **Person**, **Project** und **Publication** stellen administrative Metadaten dar, die die Beschreibungskomponente *Veröffentlichung* berücksichtigen.

Die so modellierten Korpusmetadaten sind damit tief strukturiert, extensiv, einheitlich, technisch-abstrakt und produktorientiert entwickelt. Die Definition von Korpus, die hier erarbeitet ist, stützt sich auf ein Set an abstrakten Korpuseigenschaften, das aus einer Menge an historischen Korpora gewonnen wurde. Damit kann das MKM als eine Art Merkmalsspezifikation von Korpora verstanden werden, die nicht auf der Grundlage von fachbezogenen Konzepten fußt, sondern auf gemeinsamen technisch-abstrakten Eigenschaften basiert. So leistet das MKM einen definitorischen Beitrag in der korpusbasierten Forschung.

Realisierung des MKM und Anwendung Dieses Metamodell wird mit Hilfe der TEI und ihrer Modellierungssprache durch die ODD realisiert (Kapitel 7). Dieser Ansatz ist insofern folgerichtig, da die TEI im Vergleich zu anderen Ansätzen eine hohe überfachliche Akzeptanz besitzt und das TEI-Framework Anpassungen beziehungsweise Spezifikationen nachvollziehbar ermöglicht. Die TEI-Spezifikationen erlauben ebenfalls eine tief strukturierte Realisierung und mehrere Validierungsmechanismen, die für eine Anwendung notwendig sind. Für jede der zentralen Klassen **Corpus**, **Document** und **Annotation** wird eine ODD erstellt. Die TEI-MKM-Metadaten zu einer Annotation, zu einem Dokument oder Korpus werden dann jeweils im `teiheader` abgelegt. Durch eine TEI-Spezifikation, die als originäres Subset der aktuellen Version P5 entwickelt ist, existiert eine Chance auf Interoperabilität zu anderen Ansätzen. Dieser Ansatz ist dennoch innovativ, da in den hier entwickelten TEI-Spezifikationen der Metadatenbezug von einem Dokument in der TEI-konformen Datei auf ein Kor-

pus außerhalb der TEI-konformen Datei ausgeweitet wurde. Mit diesem neuen Bezug wird TEI-XML als reines Metadatenformat genutzt.

Die Realisierung des MKM kann in Anwendungen, die die Wiederverwendung von Korpora unterstützen, angewendet werden. So kann in einem Repositorium wie LAUDATIO auf Basis der technisch-abstrakten produktorientierten Metadaten für jedes gespeicherte Korpus eine einheitliche, ausführliche Korpusdokumentation angezeigt werden sowie auf Basis dieser Metadaten eine Metadatenfreitext- und Facettensuche aufgebaut werden. Damit sind dann auch verschiedene Handlungen auf Basis der Metadaten innerhalb einer Anwendung möglich. Zusammenfassend werden mit dieser Arbeit die Voraussetzungen für die Wiederverwendung von historischen Korpora geschaffen und damit wird auch dazu beigetragen, die Nachhaltigkeit von historischen Korpora zu fördern.

9 Diskussion und Ausblick

Der hier vorgeschlagene Ansatz versucht mit Hilfe eines Metamodells für Korpusmetadaten und seiner Realisierung als Grundlage für eine umfangreiche, einheitliche und tief strukturierten Korpusdokumentation, die Wiederverwendung von historischen Korpora zu unterstützen. Dies kann nur in Verbindung mit einem freien Zugang zu den Korpusdaten gelingen. Korpuserstellerinnen und -ersteller von historischen Korpora müssen die Möglichkeit haben, ihre Korpora über Repositorien wie LAUDATIO frei zu veröffentlichen, sodass diese Daten geteilt werden können. Das Teilen von Korpora ist im gleichem Maße Voraussetzung für deren Wiederverwendung wie die Korpusdokumentation selbst. Dem Teilen geht eine Veröffentlichung der Daten voraus, die inklusive einer Dokumentation einen wesentlichen Teil des Datenmanagements darstellt.

Nachhaltigkeit und Qualität Wie bereits in Abschnitt 4.7 angeführt, ginge eine tiefere Diskussion des Qualitätsbegriffs und die Überprüfung von Qualität und Qualitätsmerkmalen von Metadaten über den Rahmen dieser Arbeit weit hinaus. In dieser Arbeit wird daher ganz allgemein davon ausgegangen, dass Forschungsdaten dann nachhaltig sind, wenn sie auch wiederverwendet werden (können) (vgl. Jensen et al. 2011; Simons und Bird 2008). Wie Simons und Bird (2008) treffend feststellen, wird eine Ressource genutzt, wenn sie existiert und sie nutzbar und relevant für die eigene Forschung ist. Repositorien stellen die nachhaltige Speicherung der Korpora und damit deren Existenz sicher. Durch eine umfangreiche einheitliche Dokumentation, wie sie hier mit dem MKM vorgeschlagen wird, kann die Auffindbarkeit und Nutzbarkeit von historischen Korpora gefördert werden. Die Metadaten, die das Korpus als Ergebnis (Produkt) eines Forschungsdatenzyklus beschreiben, helfen Nutzerinnen und Nutzern dabei, die einzelnen Erstellungs- und Bearbeitungsschritte einheitlich für viele Korpora nachvollziehen zu können. Weiterhin unterstützen die tief strukturierten, deskriptiven Metadaten zu den Annotationen und zu den historischen Vorlagen Nutzerinnen und Nutzer darin, zu bewerten, welche Ressource für ihre Forschung geeignet ist.

Einen weiteren Rahmen für die Definition und Prüfung der Nachhaltigkeit von Forschungsdaten und Anwendungen stellen beispielsweise das Data Seal of Approval[182] oder die FAIR Guiding Principles for scientific data management and stewardship (Wilkinson et al. 2016). (Wilkinson et al. 2016) wollen mit diesen Prinzipien die Wiederverwendung von Forschungsdaten und die umfassende Beschreibung mit Metadaten unterstützen. Mit dem Data Seal of Approval werden Repositorien für Forschungsdaten hinsichtlich Kriterien der Nachhaltigkeit geprüft und ggf. ausgezeichnet. In einem weiteren Schritt kann das MKM also beispielsweise im Rahmen der FAIR Guiding Principles diskutiert werden. Zusammen mit der Anwendung in dem LAUDATIO-Repositoriums kann eine Evaluation im Rahmen des Data Seal of Approval erfolgen.

Methodik Die Entwicklung eines Metamodells für Korpusmetadaten setzt ein abstraktes Verständnis über die Korpustypen und deren Besonderheiten voraus, die bei jeder Art der Wiederverwendung auch Nutzerinnen und Nutzer berücksichtigen müssen (Kapitel 2). Dieser Ansatz ist unabhängig von Formaten und konkreten Annotationsmodellen. Weiterhin besitzt dieses Metamodell einen kleineren Bezugsrahmen als beispielsweise der Metadatenstandard DC und legt einen über die Ressource begrenzten Anwendungsrahmen vergleichbar mit den Ansätzen der TEI, IMDI und des METS fest, die ebenfalls jeweils auf bestimmte Ressourcentypen (andere als historische Korpora) spezialisiert sind (Kapitel 5). Dabei sind die Fragen des Zwecks, der Granularität und der Bezugsgrößen zentral, die mit dem Metamodell beantwortet werden. Welche Eigenschaften welcher Daten sollen für welchen Zweck beschrieben werden? Sollen nur wenige, allgemeine Metadaten zu vielen unterschiedlichen Daten angegeben werden? Wenn sich die zu beschreibende Menge an Objekten wenige Eigenschaften teilt, dann können nur wenige allgemeine Metadaten allen Objekten einheitlich zugewiesen werden. Zusammengefasst heißt das: Je mehr Eigenschaften sich die zu beschreibende Menge an Objekten teilt, desto detaillierter und umfangreicher können einheitliche Metadaten vergeben werden. Je höher der Spezialisierungsgrad auf eine Ressource ist, desto umfangreicher können Metadaten auch in Bezug auf einen bestimmten Zweck wie beispielsweise Wiederverwendungsszenarien eingesetzt werden. In jedem Fall ist eine klare Definition des Zwecks der Modellierung notwendig, um eine Auswahl der zu beschreibenden Merkmale zu motivieren (Kapitel 3 und Abschnitt 5.1). Das Metamodell erfasst somit die zu beschreibenden Typen von Objekten sowie deren für den Zweck relevanten Merkmale und ermög-

[182] `http://www.datasealofapproval.org` (besucht am 27.01.2017).

licht eine formale, abstrakte, tief strukturierte und formatunabhängige Beschreibung verschiedenster historischer Korpora. Die Instanziierungen des Metamodells können dann als architektonische Grundlage für eine Metadatensuche in einem Repositorium eingesetzt werden und ermöglichen damit tief strukturierte, einheitliche Suchen nach verschiedensten Eigenschaften historischer Korpora.

Dem gegenüber steht der Ansatz der CMDI, der jeder Nutzerin oder jedem Nutzer ermöglicht, ein eigenes, unabhängiges Metadatenset für alle möglichen Objekte zu entwickeln, ohne bestimmte Guidelines oder Richtlinien vorzugeben. Eine einheitliche Modellebene und ein Modellierungsmechanismus sind in CMDI daher nicht vorgesehen. Dies führt nun zu einer unkontrollierten Nutzung inklusive der Erstellung von Dubletten (T. Eckart 2016: 165). Dies geschieht auch aus den in Kapitel 2 beschriebenen Gründen und ist daher wenig überraschend. Jede Nutzerin oder jeder Nutzer verwendet dabei eigene Modelle der Objekte, die er oder sie beschreiben will. Einerseits möchte die CMDI den Nutzerinnen und Nutzern größtmögliche Freiheit lassen, anderseits wird seitens der CMDI dieser „Wildwuchs“, gemeint sind nutzerspezifische Metadatenschemata, als problematisch erkannt. Um diese verschiedenen Metadatenschemata für eine Metadatensuche über alle Metadaten wieder aufeinander abbilden zu können, wie es T. Eckart (2016) diskutiert, bedarf es einer Deutungshoheit über die einzelnen Nutzerschemata hinweg, wie sie CMDI beansprucht. Mappings der diversen Profile (oder ihrer Komponenten) werden durchgeführt, ohne aber deren jeweiligen spezifischen Kontext mit den expliziten oder impliziten Modellen und deren jeweiligen Beschreibungszweck der Nutzerinnen und Nutzer genau kennen zu können.

Die methodische Frage, die sich daraus ergibt und sehr unterschiedlich zwischen den verschiedenen Ansätzen diskutiert wird, ist, inwieweit eine Post-hoc-Zusammenfassung von Metadaten, deren jeweilige Beschreibungsmodelle nicht im Einzelnen nachvollzogen oder komplett aufeinander abgebildet werden können, vorteiliger ist als ein modellbasierter Ansatz zur Erstellung von Metadaten, der den Nutzerinnen und Nutzern einen Objektbezug, eine klare Informationsarchitektur, Guidelines sowie einen Spielraum für die Realisierung von Metadaten vorgibt.

Die hier vorliegende Arbeit hat einen Ansatz gewählt, der die Menge der zu beschreibenden Objekte auf den Korpustyp historisches Textkorpus eingrenzt und der sich auf einen bestimmten Zweck (Wiederverwendung) konzentriert. Für diesen Korpustyp existiert noch kein entsprechendes Metadatenschema, das alle Anforderungen erfüllen kann, die der Zweck stellt. Vielmehr fehlt ein Konzept, wie die unterschiedlichen Informationen, die für eine solche Korpusdokumentation benötigt werden, mit

Hilfe von Metadaten strukturiert umgesetzt werden können. Dafür wird ein Metamodell für Korpusmetadaten vorgeschlagen, das in einem ersten Schritt über die vielen Eigenschaften von historischen Korpora abstrahiert, die relevant für eine Korpusdokumentation zum Zweck der Wiederverwendung sind. Ein solche Vorgehensweise stützt sich auf die konkrete Datenlage und deren Wiederverwendungsszenarien. Die verschiedenen Nutzergruppen haben einen gemeinsamen Bezugspunkt: die Forschungsdaten und ihre Verwendung als empirische Forschungsgrundlage. Diese historischen Korpora unterscheiden sich zwar stark in vielen Aspekten, es können aber auch Gemeinsamkeiten herausgearbeitet werden, die von allgemeinen Prinzipien wie dem Forschungsprozess und dem Forschungsdatenzyklus sowie der Datenarchitektur von historischen Korpora abgeleitet sind. Ergebnis dieser hier vorliegenden Arbeit ist ein einheitliches, tief strukturiertes und umfassendes Metamodell für Korpusmetadaten. Wenn Korpora überfachlich von verschiedenen Nutzerinnen imd Nutzern ertellerunabhängig erschlossen werden sollen, dann ist eine einheitliche Beschreibung und eine Suche über diese Beschreibungen sinnvoll und schafft Vergleichbarkeit zwischen den verschiedenen Korpora, die bereits bei der korpuseigene Dokumentation hergestellt wird und nicht post hoc erzeugt werden muss.

TEI-Konformität und Anbindung an andere Standards Die Realisierung des Metamodells für Korpusmetadaten in TEI ist eine relativ innovative Nutzung der TEI als reines Metadatenmodell. Die bisherige Realisierung des MKM weitet den Bezugsrahmen der TEI-Metadaten im `teiheader` auf TEI- und datenexterne Objekte aus. Der `text`-Bereich der TEI-konformen Datei bleibt hingegen leer. In einem nächsten Schritt kann geprüft werden, ob und wie einige Objekte, die jeweils in den einzelnen TEI-Spezifikationen durch Metadaten beschrieben werden, auch in den `text`-Bereich der TEI-konformen Dateien integriert werden können. Damit würde das TEI-Prinzip, innerhalb einer Datei Metadaten und das durch Metadaten Beschriebene zu realisieren, wieder aufgenommen werden können (Abschnitt 5.5). Dazu müssen Überlegungen angestellt werden, wie die für historische Korpora beschriebenen Korpusarchitekturen in die TEI-Welt integriert werden können. Einen solchen Ansatz diskutieren beispielsweise Bański et al. (2016) mit den eingebetteten Standoff-Annotationen in derTEI.

Darüber hinaus kann der Anwendungsbereich des MKM vergrößert werden, indem es auch mit anderen Metadatenframeworks wie DC oder CMDI realisiert wird. Eine Realisierung mit DC würde zwar ein hohen Informationsverlust bedeuten, aber die Sichtbarkeit von Korpora in anderen Anwendungen, die diesen Standard unter-

stützen, erhöhen. Eine Realisierung mit anderen Metadatenstandards ist möglich, da das MKM formatunabhängig mit Hilfe der UML entwickelt ist.

Erweiterung des MKM Eine der naheliegenden nächsten Schritte ist, den Rahmen und den Bezug dieser Arbeit auszuweiten und zu prüfen, inwieweit das MKM auf andere Korpustypen wie moderne textbasierte Korpora oder Korpora der gesprochenen Sprache angewendet werden kann. Eine Erweiterung auf allgemein textbasierte Korpora ist bereits in dem Metamodell angelegt, weil historische Korpora einen Subtyp von textbasierten Korpora darstellen. Viele der modellierten korpuseigenen Eigenschaften treffen nicht allein auf historische Korpora zu, sondern auf alle textbasierten Korpora. Korpora aus modernen Zeitungstexten oder Briefen tragen abstrakt gesehen die gleichen Eigenschaften wie historische Korpora aus Zeitungstexten oder Briefen. Beide Typen können verschiedene Annotationskonzepte wie Spannen- oder Baumannotationen beinhalten oder mit gleichen Bearbeitungsschritten erstellt werden. Die **Document**-Metadaten ähneln sich in diesem Fällen ebenfalls, da beide vergleichbare Vorlagen besitzen, die mit bibliographischen Angaben beschrieben werden können. Modernen Texten fehlen häufiger die umfangreicheren Metadaten zu Handschrift und Niederschrift, wie sie historische Texte besitzen können. In diesem Fall wären wenige Anpassungen im MKM zu erwarten.

Bei anderen Korpustypen, wie Korpora der gesprochene Sprache, die auf einem unterschiedlichen Datentyp basieren können, stellen sich mehrere Fragen, die eine Anpassung oder Spezialisierung des MKM nach sich ziehen können. Dieselben Prinzipien von korpuseigenen Eigenschaften, die beispielsweise die Korpusarchitektur betreffen, können ebenfalls auf nicht typische Vertreter textbasierter Korpora wie Korpora mit Transkripten von gesprochener Sprache angewendet werden. So ist zu prüfen, inwieweit korpusexterne Eigenschaften wie z. B. Eigenschaften der SprecherInnen und des gesprochenen sprachlichen Materials im MKM abgebildet werden können und müssen. Dies könnte eine Erweiterung der Klasse **Person** um weitere Attribute wie *Alter* oder *Geschlecht* bedeuten. Weiterhin muss geprüft werden, inwieweit die Beziehung zu Audiosignal und Transkript im MKM abgebildet werden kann. Hierzu muss das MKM mit einer umfassenden Menge von Korpora der gesprochene Sprache abgeglichen werden. Die Ausweitung des Objektbezugs des MKM zieht dann eine Anpassungen der Realisierung in TEI mit Hilfe des ODD-Mechanismus nach sich. Da die TEI ebenfalls für Korpora der gesprochenen Sprache eingesetzt wird (Schmidt 2011), kann auf Vorarbeiten zurückgegriffen werden.

Die in dieser Arbeit entwickelten Wiederverwendungsszenarien und die Metamo-

dellierung von Akteurinnen und Akteuren stellen den Anspruch an einen – unter Vorbehalt von Irrtum und Auslassung – so umfangreichen Geltungsbereich, dass sie ohne Modifizierungen auf andere Korpustypen abgebildet werden können. Korpora können ganz allgemein mit weiterem sprachlichen Material – abgebildet mit der Klasse **Document** – oder weiteren Annotationen – abgebildet mit der Klasse **Annotation** – angereichert oder um diese reduziert sowie weiter analysiert oder konvertiert werden. Auch die Unterscheidungen der Akteurinnen und Akteure in Bezug auf ihren Kenntnisstand des Korpus zum Zeitpunkt der Wiederverwendung lassen sich auf andere Korpustypen anwenden.

Ein weiterer nächster Schritt in der (Weiter-)Entwicklung des MKM ist es, die bislang sehr allgemeinen *ID*s mit Werttyp *String* näher zu spezifizieren und deren Einsatz im MKM auszuweiten. Eine Möglichkeit, die Aussagekraft der Metadaten durch die Vergabe von *ID*s zu erhöhen, ist die Integration von Informationen aus anderen Wissensdatenbanken wie die Gemeinsame Normdatei (GND)[183], die mittels eindeutiger Referenzen entweder korpusverantwortliche Personen oder Autoren von Texterzeugnissen identifizieren können. Hier müssten die jeweiligen *ID*s im MKM, die momentan als ein sehr allgemeiner Fall modelliert sind, spezialisiert werden. Eine solche Erweiterung des MKM würde eine stärkere Referenzierungsfunktion der im MKM modellierten Metadaten innerhalb einer Korpusdokumentation und über verschiedene Korpusdokumentationen hinweg bedeuten.

Das MKM und seine Realisierung in TEI müssen qua Modellierung auch für weitere, bislang ungesehene historische Korpora einsetzbar sein. Die Aufnahme weiterer historischer Korpora wird dies zeigen. Die Nutzerfreundlichkeit und Funktionalität der Umsetzung der TEI-MKM-Metadaten in LAUDATIO zu prüfen und umfangreicher in die Informationsarchitektur der Anwendung einzubauen, stellt ein weiteres Desiderat dar. Durch Nutzerfeedback können Rückschlüsse auf noch offene Aspekte im MKM gezogen werden. Auch der Einsatz des MKM in Verbindung mit anderen Anwendungen wie Annotationstools oder Suchtools kann überlegt werden, da ein Teil der Metadaten auch in dieser Art der Anwendungen genutzt werden können.

[183] `http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html` (besucht am 31.10.2016).

Abkürzungsverzeichnis

ADHO The Alliance of Digital Humanities Organizations 43

ANNIS Search and Visualization in Multilayer Linguistic Corpora 45

ATILF Analyse et Traitement Informatique de la Langue Française 98

BeMaTaC Berlin Map Task Corpus 27

CC Creative Commons 58

CENDARI Collaborative European Digital Archive Infrastructure 46

CLARIN Common Language Resources and Technology Infrastructure 8

CMDI Component MetaData Infrastructure 75

COSMAS Corpus Search, Management and Analysis System 45

COW Corpora from the Web 27

CQP Corpus Query Processor 45

DANS Data Archiving and Networked Services 8

DARIAH Digital Research Infrastructure for the Arts and Humanities 8

DC Dublin Core 96

DCMES Dublin Core Metadata Element Set 97

DMCI Dublin Core Metadata Initiative 9

DCR Data Category Registry 102

DDD-AHD Deutsch Diachron Digital – Referenzkorpus Altdeutsch 29

deWaC deutsches Web als Corpus 27

DFG Deutschen Forschungsgemeinschaft 15

DGD Freiburger Korpus der Datenbank für Gesprochenes Deutsch 66

DHd Digital Humanities im deutschsprachigem Raum 43

DOBES Dokumentation Bedrohter Sprachen 100

DTA Deutsches Textarchiv 25

DTD Document Type Definition 149

EADH European Association for the Digital Humanities 43

EXMARaLDA Extensible Markup Language for Discourse Annotation 38

Falko Fehlerannotierte Lernerkorpus 27

FRBR Functional Requirements for Bibliographic Records . . 47

GECO Gesprächscorpus 27

GerManC German Manchester Corpus 10

GND Gemeinsame Normdatei 188

HIPKON Historisches Predigtenkorpus zum Nachfeld 53

HiTS Historisches Tagset 29

HZSK Hamburger Zentrum für Sprachkorpora 46

ICLTT Institute for Corpus Linguistics and Text Technology 148

IMDI ISLE Meta Data Initiative 87

ISO International Organization for Standardization 138

ISOcat ISO Catalogue 102

KAJUK Kasseler Junktionskorpus 56

LAUDATIO Long-term Access and Usage of Deeply Annotated Information 45

LDC Linguistic Data Consortium 9

MARC MAchine-Readable Cataloging 107

MKM Metamodell für Korpusmetadaten 19

MEI Music Encoding Initiative 108

METS Metadata Encoding and Transmission Standard 17

MODS Metadata Object Description Schema 107

NISO National Information Standards Organization 172

NLP Natural Language Processing 44

ODD TEI-Spezifikation One Document does it all 108

OLAC Open Language Archives Community 9

OPAC Online Public Access Catalogue 17

PAULA Potsdamer AUstauschformat Linguistischer Annotationen 40

PID Persistent Identifier 102

pos Part of Speech 30

PTB Penn Treebank 76

RDA Research Data Alliance 9

RDF Resource Description Framework 85

RELAX NG REgular LAnguage for XML Next Generation 149

RIDGES Register in Diachronic German Science 5

SGML Standard Generalized Markup Language 56

STTS Stuttgart-Tübingen-Tagset 29

TCF Text Corpus Format 39

TEI Text Encoding Initiative 6

UIMA Unstructured Information Management Application . . 75

UML Unified Modeling Language 6

VLO Virtual Language Observatory 101

XML Extensible Markup Language 33

Abbildungsverzeichnis

Referenzen

Abney, Steven P. und Bird, Steven (2011): Towards a data model for the Universal Corpus. In: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*. 120–127.

Ágel, Vilmos und Hennig, Mathilde (2006): Theorie des Nähe- und Distanzsprechens. In: Ágel, Vilmos und Hennig, Mathilde (Hrsg.): *Grammatik aus Nähe und Distanz*. Tübingen: Niemeyer. 3–31.

Ágel, Vilmos und Hennig, Mathilde (2008): *KAJUK (Version 1.1)*. Universität Kassel. Justus-Liebig-Universität Gießen, Kasseler Junktionskorpus. URL: `http://hdl.handle.net/11022/0000-0000-2102-8`.

Allwood, Jens (2008): Multimodal corpora. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics*. Berlin: De Gruyter. 207–224.

Apollon, Daniel; Bélisle, Claire und Régnier, Philippe, Hrsg. (2014): *Digital Critical Editions*. Urbana, Chicago und Springfield: University of Illinois Press.

Archer, Dawn; Kytö, Merja; Baron, Alistair und Rayson, Paul (2015): Guidelines for normalising Early Modern English corpora: Decisions and justifications. In: *ICAME Journal*(39). 5–24.

Atwell, Eric (2008): Development of tag sets for part-of-speech tagging. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics*. Berlin: De Gruyter. 501–527.

Baca, Murtha, Hrsg. (2008): *Introduction to Metadata: pathways to digital information*. 2. Auflage. Los Angeles: Getty Research Insitute.

Baillot, Anne und Seifert, Sabine (2013): The Project "Berlin Intellectuals 1800–1830" between Research and Teaching. In: *Journal of the Text Encoding Initiative* 4. DOI: `10.4000/jtei.707`.

Ballier, Nicolas und Martin, Philippe (2015): Speech annotation of learner corpora. In: Granger, Sylviane; Gilquin, Gaëtanelle und Meunier, Fanny (Hrsg.): *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge Univ. Press. 107–134.

Bański, Piotr; Gaiffe, Bertrand; Lopez, Patrice; Meoni, Simon; Romary, Laurent; Schmidt, Thomas und Stadler, Peter, Witt, Andreas (2016): *Wake up, standOf!* TEI Conference 2016. Wien. URL: http://tei2016.acdh.oeaw.ac.at.

Bański, Piotr und Przepiórkowski, Adam (2009): Stand-off TEI Annotation: the Case of the National Corpus of Polish. In: *Proceedings of the Third Linguistic Annotation Workshop*. 64–67.

Baron, Alistair und Rayson, Paul (2008): VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In: *Proceedings of Postgraduate Conference in Corpus Linguistics*. URL: http://acorn.aston.ac.uk/conf_proceedings.html (besucht am 19. 12. 2016).

Baroni, Marco; Bernardini, Silvia; Ferraresi, Adriano und Zanchetta, Eros (2009): The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. In: *Language Resources and Evaluation* 43(3). 209–226. DOI: 10.1007/s10579-009-9081-4.

Barteld, Fabian; Schröder, Ingrid und Zinsmeister, Heike (2016): Dealing with word-internal modification and spelling variation in data-driven lemmatization. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 52–62.

Bartsch, Nina; Dipper, Stefanie; Eschke, Lars; Herbers, Birgit; Klein, Thomas; Kwekkeboom, Sarah; Weber, Elke und Wegera, Klaus-Peter (2011): *Annotiertes Referenzkorpus Mittelhochdeutsch (1050–1350)*. Göttingen. URL: http://www.sfs.uni-tuebingen.de/~versley/dgfs-poster-2011/poster_03.pdf (besucht am 19. 12. 2016).

Belz, Malte (2013): Disfluencies und Reparaturen bei Muttersprachlern und Lernern: eine kontrastive Analyse. Diss. Berlin: Humboldt-Universität zu Berlin. URL: urn:nbn:de:kobv:11-100215407 (besucht am 19. 12. 2016).

Belz, Malte; Odebrecht, Carolin; Perlitz, Laura; Schnelle, Gohar und Voigt, Vivian (2016): *Annotationsrichtlinien zu Ridges Herbology Version 5.0*. Berlin. URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/ridgesv5_2016-10-19.pdf (besucht am 10. 11. 2017).

Bennett, Paul; Durrell, Martin; Ensslin, Astrid; Scheible, Silke und Whitt, Richard; (2007): *GerManC (Version 1.0)*. University of Manchester. German Manchester Corpus Project. URL: http://hdl.handle.net/11022/0000-0000-2D1B-1.

Berman, Fran; Wilkinson, Ross und Wood, John (2014): Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance. In: *D-Lib Magazine* 20(2). DOI: 10.1045/january2014-berman.

Biber, Douglas (1993): Representativeness in Corpus Design. In: *Literary and Linguistic Computing*(8). 243–257.

Biber, Douglas und Conrad, Susan (1999): Lexical Bundles in Conversation and Academic Prose. In: Hasselgard, Hilde und Oksefjell, Signe (Hrsg.): *Out of Corpora*. Amsterdam: Radopi. 181–190.

Biber, Douglas und Conrad, Susan (2009): *Register, Genre, and Style*. (= Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Bird, Steven und Liberman, Mark (2000): A Formal Framework for Linguistic Annotation. In: *CoRR*(cs.CL/0010033). 1–29. URL: http://arxiv.org/abs/cs.CL/0010033 (besucht am 24. 08. 2016).

Bird, Steven und Simons, Gary (2001): The OLAC Metadata Set and Controlled Vocabularies. In: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 7–18. URL: http://arxiv.org/abs/cs/0105030 (besucht am 30. 01. 2016).

Bird, Steven und Simons, Gary (2003): Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. In: *Computers and the Humanities* 37(4). 375–388. URL: http://www.jstor.org/stable/30204912 (besucht am 16. 09. 2016).

Bodard, Gabriel (2010): EpiDoc: Epigraphic documents in XML for publication and interchange. In: Feraudi-Gruénais, Francisca (Hrsg.): *Latin on Stone*. Lanham, MD: Lexington Books. 101–117.

Bollmann, Marcel; Dipper, Stefanie; Krasselt, Julia und Petran, Florian (2012): Manual and semi-automatic normalization of historical spelling: case studies from Early New High German. In: *Proceedings of KONVENS 2012*. 342–350. URL: http://www.oegai.at/konvens2012/proceedings/51_bollmann12w/ (besucht am 24. 08. 2016).

Borgmann, Christine L. (2012): The conundrum of sharing research data. In: *Journal of the American Society for Information Science and Technology* 63(6). 1059–1087. DOI: 10.2139/ssrn.1869155.

Bosch, Thomas und Eckert, Kai (2014): Requirements on RDF Constraint Formulation and Validation. In: DCMI (Hrsg.): *DC 2014*. 95–108.

Box, George E. P. (1979): Robustness in the strategy of scientific model building. In: Launer, Robert L. und Wilkinson, Graham N (Hrsg.): *Robustness in Statistics*. New York, San Francisco und London: Academic Press. 201–236.

Broeder, Daan; Declerck, Thierry; Romary, Laurent; Choukri, Khalid; Uneson, Markus; Strömqvist, Sven und Wittenburg, Peter (2004): A large Metadata Domain for Language Ressources. In: *Proceedings of the LREC*. 369–372.

Broeder, Daan; Kemps-Snijders, Marc; Van Uytvanck, Dieter; Windhouwer, Menzo; Withers, Peter; Wittenburg, Peter und Zinn, Claus (2010): A data category registry- and component-based metadata framework. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 43–47.

Broeder, Daan; Schuurmann, Ineke und Windhouwer, Menzo (2014): Experiences with the ISOcat Data Category Registry. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA). 4565–4568.

Broeder, Daan; Wittenburg, Peter; Declerck, Thierry und Romary, Laurent (2002): LREP: A Language Repository Exchange Protocol. In: *Third International Conference on Language Resources and Evaluation*. 1302–1305.

Buddenbohm, Stefan; Engelhardt, Claudia und Wuttke, Ulrike (2016): Angebotsgenese für ein geisteswissenschaftliches Forschungsdatenzentrum. In: *Zeitschrift für digitale Geisteswissenschaften*. DOI: 10.17175/2016_003.

Budin, Gerhard; Kabas, Heinrich und Mörth, Karlheinz (2012): Towards Finer Granularity in Metadata: Analysing the Contents of Digitised Periodicals. In: *Journal of the Text Encoding Initiative*(2). DOI: 10.4000/jtei.416.

Burnard, Lou (2013): The Evolution of the Text Encoding Initiative: From Research Project to Research Infrastructure. In: *Journal of the Text Encoding Initiative*(5). 1–13. DOI: 10.4000/jtei.811.

Burnard, Lou und Rahtz, Sebastian (2004): *RelaxNG with Son of ODD*. Montréal. URL: http://www.tei-c.org/cms/Talks/extreme2004/paper.html (besucht am 08. 01. 2017).

Burr, Elisabeth; Burkhardt, Julia; Potapenko, Elena; Sierig, Rebecca und Concepción Durán, Arámis (2015): *Das Duisburg-Leipzig Korpus romanischer Zeitungssprachen und sein Textmodell*. Graz. URL: https://dhd2015.uni-graz.at/de/nachlese/book-of-abstracts/.

Büttner, Stephan; Hobohm, Hans-Christoph und Müller, Lars (2011): Research Data Management. In: Büttner, Stephan; Hobohm, Hans-Christoph und Müller, Lars

(Hrsg.): *Handbuch Forschungsdatenmanagement.* Bad Honnef: Bock + Herchen. 13–23.

Carstensen, Kai-Uwe; Ebert, Christian; Ebert, Cornelia, Jekat, Susanne, Klabunde, Ralf und Langer, Hagen (2010): *Computerlinguistik und Sprachtechnologie: Eine Einführung.* 3. Auflage. Heidelberg: Spektrum Akademischer Verlag.

Cartwright, Nancy (1983): *How the Laws of Physics Lie.* Oxford: Clarendon Press.

Chen, Danqi und Manning, Christopher D. (2014): A Fast and Accurate Dependency Parser using Neural Networks. In: *Empirical Methods in Natural Language Processing (EMNLP).*

Cieri, Christopher und Liberman, Mark (2000): Issues in Corpus Creation and Distribution. In: *Proceedings of the Second International Conference on Language Resources and Engineering.* Athen.

Claridge, Claudia (2008): Historical Corpora. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics.* Berlin: De Gruyter. 242–259.

CLARIN-D AP 5 (2012): *CLARIN-D User Guide.* URL: `http://media.dwds.de/clarin/userguide/userguide-1.0.1.pdf` (besucht am 05. 01. 2017).

Coniglio, Marco; Donhauser, Karin und Schlachter, Eva (2014): *HIPKON: Historisches Predigtenkorpus zum Nachfeld (Version 1.0).* Humboldt-Universität zu Berlin. SFB 632 Teilprojekt B4. URL: `http://hdl.handle.net/11022/0000-0000-2D18-4`.

Coniglio, Marco; Donhauser, Karin; Schlachter, Eva; Rasskazova, Oxana; Odebrecht, Carolin; Wirth, Matthias und Miltenberger, Anke (2016): *Historisches Predigtenkorpus zum Nachfeld (HIPKON Version 1.0) - Dokumentation.* DOI: `10.18452/13681`.

Conrad, Susan (1996): Investigating academic texts with corpus-based techniques: An example from biology. In: *Linguistics and Education* Volume 8(3). 299–326.

Cover, Robin C. und Robinson, Peter M.W. (1995): Encoding Textual Criticism. In: Ide, Nancy und Véronis, Jean (Hrsg.): *Text Encoding Initiative.* (= Computers and the Humanities, Bd. 29, Nr. 1, 2 & 3). Dorbrecht [u.a]: Kluwer Academic Publishers. 123–136.

Coyle, Karen (2005): Understanding Metadata and its Purpose. In: *Journal of Academic Librarianship* 2. 160–163.

Demske, Ulrike (2005): *Mercurius-Baumbank (Version 1.1).* Universität Potsdam. Mercurius Projekt. URL: `http://hdl.handle.net/11022/0000-0000-467D-6`.

Demske, Ulrike (2007): Das Mercurius-Projekt: eine Baumbank für das Frühneuhochdeutsche. In: Zifonun, Gisela und Kallmeyer, Werner (Hrsg.): *Sprachkorpora.*

(= Jahrbuch des Instituts für deutsche Sprache 2006). Berlin: De Gruyter. 91–104.

Deutsche Forschungsgemeinschaft (2009): *Empfehlungen zur gesicherten Aufbewahrung und Bereitstellung digitaler Forschungsprimärdaten*. Hrsg. von Ausschuss für Wissenschaftliche Bibliotheken und Informationssysteme. Bonn. URL: `http://www.dfg.de/download/pdf/foerderung/programme/lis/ua_inf_empfehlungen_200901.pdf` (besucht am 13. 01. 2016).

Dietterle, Burkhard (2016): *Binnenklammern im deWaC: Kolloquium Korpuslinguistik*. Berlin. URL: `https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/lehre/wise-1516/ko_5220053` (besucht am 01. 08. 2016).

Digital Curation Centre (2010): *DCC Curation Lifecycle Model*. URL: `http://www.dcc.ac.uk/resources/curation-lifecycle-model` (besucht am 29. 02. 2016).

DiPersio, Denise; Cieri, Christopher und Jaquette, Daniel (2016): Data Management Plans and Data Centers. In: *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*. 2496–2501.

Dipper, Stefanie (2005): XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: *Proceedings of Berliner XML Tage 2005 (BXML 2005)*. Berlin. 39–50.

Dipper, Stefanie; Donhauser, Karin; Klein, Thomas; Linde, Sonja; Müller, Stefan und Wegera, Klaus-Peter (2013): HiTS: Ein Tagset für historische Sprachstufen des Deutschen. In: Zinsmeister, Heike; Heid, Ulrich und Beck, Kathrin (Hrsg.): *Das Stuttgart-Tübingen Wortarten-Tagset*. (= Journal for Language Technology and Computational Linguistics, Bd. 28(1)). 85–137.

Dipper, Stefanie; Götze, Michael und Stede, Manfred (2004): Simple Annotation Tools for Complex Annotation Tasks: An Evaluation. In: *Proceedings of the LREC*. 54–62.

Dipper, Stefanie; Krasselt, Julia und Schultz-Balluff, Simone (2015): Creating synopses of 'parallel' historical manuscripts and early prints. Alignment guidelines, evaluation, and applications. In: Gippert, Jost und Gehrke, Ralf (Hrsg.): *Historical Corpora*. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd. 5). Tübingen: Narr. 137–150.

Dipper, Stefanie und Schultz-Balluff, Simone (2013): The Anselm Corpus: Methods and Perspectives of a Parallel Aligned Corpus. In: *Proceedings of the NODALIDA Workshop on Computational Historical Linguistics*. 27–42.

Donahue, Christiane und Lillis, Theresa (2014): Models of Writing and Text Production. In: Eva-Maria Jakobs und Daniel Perrin (Hrsg.): *Handbook of Writing and Text Production*. Berlin: De Gruyter. 55–78.

Donhauser, Karin (2015): Das Referenzkorpus Altdeutsch: Das Konzept, die Realisierung und die neuen Möglichkeiten. In: Gippert, Jost und Gehrke, Ralf (Hrsg.): *Historical Corpora*. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd. 5). Tübingen: Narr. 35–49.

Donhauser, Karin; Gippert, Jost und Lühr, Rosemarie (2014): *Referenzkorpus Altdeutsch(Version 0.1)*. Humboldt-Universität zu Berlin, Friedrich-Schiller-Universität Jena, Goethe Universität Frankfurt. Deutsch Diachron Digital. URL: `http://hdl.handle.net/11022/0000-0000-7FC2-7`.

Dreyer, Malte und Vollmer, Andreas (2016): *An Integral Approach to Support Research Data Management at the Humboldt-Universität zu Berlin*. Thessaloniki. DOI: `10.13140/RG.2.1.2623.8961`.

Druskat, Stephan; Bierkandt, Lennart; Gast, Volker; Rzymski, Christoph und Zipser, Florian (2014): Atomic: an open-source software platform for multi-layer corpus annotation. In: Ruppert, Josef und Faaß, Gertrud (Hrsg.): *Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache*. 228–234. URL: `http://hildok.bsz-bw.de/frontdoor/index/index/docId/266` (besucht am 08. 10. 2016).

Dudenredaktion (2016): *Duden Online Wörterbuch*. Hrsg. von Bibliographisches Institut GmbH. Berlin. URL: `http://www.duden.de/woerterbuch` (besucht am 03. 08. 2016).

Dumont, Stefan (2016): Fürstinnenkorrespondenzen: Nachnutzung eines Briefkorpus aus LAUDATIO. In: *Workshop Entwicklung und Nutzung interdisziplinärer Repositorien für historische textbasierte Korpora. Digital Humanities im deutschsprachigem Raum*. Leipzig.

Durrell, Martin; Ensslin, Astrid und Bennett, Paul (2007): The GerManC project. In: *Sprache und Datenverarbeitung*(31). 71–80.

Eckart, Kerstin (2015): *Nachhaltige Ressourcen durch Dokumentation von Verarbeitungsschritten: Metadaten, Prozessmetadaten und Workflowvisualisierung: Kolloquium Korpuslinguistik*. Berlin.

Eckart, Thomas (2016): Einsatz und Bewertung komponentenbasierter Metadaten in einer föderierten Infrastruktur für Sprachressourcen am Beispiel der CMDI. Diss. Leipzig: Universität Leipzig.

Erdmann, Oskar, Hrsg. (1973): *Otfrids Evangelienbuch*. 6. Auflage besorgt von Ludwig Wolff. (= Altdeutsche Textbibliothek 49). Tübingen.

Federico, Annette (2015): *Engagements with Literature: Engagements with Close Reading*. Florence: Routledge.

Ferrucci, David und Lally, David (2004): UIMA: an architectural approach to unstructured information processing in the corporate research environment. In: *Natural Language Engineering* 10(3-4). DOI: 10.1017/S1351324904003523.

Fleischer, Jürg und Schallert, Oliver (2011): *Historische Syntax des Deutschen: Eine Einführung*. Tübingen: Narr.

Gerdes, Kim (2013): Collaborative Dependency Annotation. In: *Proceedings of the Second International Conference on Dependency Linguistics*. 88–97.

Geyken, Alexander (2013): Wege zu einem historischen Referenzkorpus des Deutschen: Das Projekt Deutsches Textarchiv. In: Hafemann, Ingelore (Hrsg.): *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. (= Thesaurus Linguae Aegyptiae, Bd. 4). 221–234.

Gilliland, Anne J. (2008): Setting the Stage. In: Baca, Murtha (Hrsg.): *Introduction to Metadata*. Los Angeles: Getty Research Insitute. 1–19.

Gippert, Jost und Gehrke, Ralf, Hrsg. (2015): *Historical Corpora*. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd. 5). Tübingen: Narr.

Gooding, Paul; Terras, Melissa und Warwick, Claire (2013): The myth of the new: Mass digitization, distant reading, and the future of the book. In: *Literary and Linguistic Computing* 28(4). 629–639. DOI: 10.1093/llc/fqt051.

Greenberg, Jane; Swauger, Shea und Feinstein, Elena M. (2013): Metadata Capital in a Data Repository. In: DCMI (Hrsg.): *DC-2013–The Lisbon Proceedings*. 140–150. URL: http://dcpapers.dublincore.org/pubs/issue/view/165 (besucht am 23. 12. 2016).

Greenstein, Daniel und Burnard, Lou (1995): Speaking with One Voice: Encoding Standards and the Prospects for an Integrated Approach to Computing in History. In: Ide, Nancy und Véronis, Jean (Hrsg.): *Text Encoding Initiative*. (= Computers and the Humanities, Bd. 29, Nr. 1, 2 & 3). Dorbrecht [u.a]: Kluwer Academic Publishers. 137–148.

Hajič, Jan and Ciaramita, Massimiliano and Johansson, Richard and Kawahara, Daisuke and Mart\'ı, Maria Ant\'onia and M\'arquez, Llu\'is and Meyers, Adam and Nivre, Joakim and Pad\'o, Sebastian and Štěp\'anek, Jan and Straň\'ak, Pavel and Surdeanu, Mihai and Xue, Nianwen and Zhang, Yi (2009): The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In:

Proceedings of the Thirteenth Conference on Computational Natural Language Learning. 1–18. URL: http://www.aclweb.org/anthology/W09-1201 (besucht am 21. 01. 2018).

Haugen, Odd Einar und Apollon, Daniel (2014): The Digital Turn in Textual Scholarship. In: Apollon, Daniel; Bélisle, Claire und Régnier, Philippe (Hrsg.): *Digital Critical Editions*. Urbana, Chicago und Springfield: University of Illinois Press.

Haynes, David (2004): *Metadata for information management and retrieval*. London: facet publishing.

Hedges, Mark; Neuroth, Heike; Smith, Kathleen M.; Blanke, Thomas; Romary, Laurent; Küster, Marc und Illingworth, Malcom (2013): TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructure for Textual Scholarship. In: *Journal of the Text Encoding Initiative*(5). 1–13.

Heid, Ulrich; Schmid, Helmut; Eckart, Kerstin und Hinrichs, Erhard W. (2010): A Corpus Representation Format for Linguistic Web Services: The D-SPIN Text Corpus Format and its Relationship with ISO Standards. In: *Proceedings of the Seventh International Conference on Language Resources and Evaluation*. 494–499.

Hennig, Mathilde, Hrsg. (2013a): *Die Ellipse: Neue Perspektiven auf ein altes Phänomen*. (= Linguistik – Impulse & Tendenzen, Bd. 52). Berlin/Boston: De Gruyter.

Hennig, Mathilde (2013b): The Kassel Corpus of Clause Linking. In: Bennett, Paul; Durrell, Martin; Scheible, Silke und Whitt, Richard J. (Hrsg.): *New Methods in Historical Corpora*. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd. 3). Tübingen: Narr. 207–219.

Herzog, Gottfried; Heid, Ulrich; Trippel, Thorsten; Bański, Piotr; Romary, Laurent; Schmidt, Thomas; Witt, Andreas und Eckart, Kerstin (2015): Recent Initiatives towards New Standards for Language Resources. In: *Proceedings of the Int. Conference of the German Society for Computational Linguistics and Language Technology*. 154–156.

Heßbrüggen-Walter, Stefan (2016): Modellierung: eine Begriffsbestimmung. In: *DHd 2016. Modellierung Vernetzung Visualisierung*. 164–166.

Hider, Philip, Hrsg. (2012): *Information Resource Description: Creating and Managing Metadata*. London: facet publishing.

Himmelmann, Nikolaus P. (2012): Linguistic Data Types and the Interface between Language Documentation and Description. In: *Language Documentation & Conservation*(6). 187–207.

Hinrichs, Erhard W. und Krauwer, Steven (2014): The CLARIN Research Infrastructure: Resources and Tools for E-Humanities Scholars. In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA). 1525–1531.

Hübner, Gert (2006): *Ältere deutsche Literatur: Eine Einführung*. Tübingen und Basel: Francke.

Humboldt-Universität zu Berlin (2014): *Grundsätze zum Umgang mit Forschungsdaten an der Humboldt-Universität zu Berlin*. URL: `https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/hu-fdt-policy/view` (besucht am 06. 06. 2016).

Hundt, Susanne (2008): Text corpora. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics*. Berlin: De Gruyter. 168–186.

Hunston, Susan (2008): Collection strategies and design decisions. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics*. Berlin: De Gruyter. 154–168.

Hunter, Jane (2003): Working Towards MetaUtopia: A Survey of Current Metadata Research. In: *Library Trends, Organizing the Internet* 52(2).

Ide, Nancy und Sudermann, Keith (2014): The Linguistic Annotation Framework: a standard for annotation interchange and merging. In: *Language Resources and Evaluation* 48(3). 395–418.

Ide, Nancy; Sudermann, Keith; Pustejovsky, James; Verhagen, Marc und Cieri, Christopher (2016): The Language Application Grid and Galaxy. In: *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference*. 457–462.

IDS, Mannheim (2013): *MaKoHiZZ (Version 1.0)*. Institut für Deutsche Sprache Mannheim.Mannheimer Korpus Historischer Zeitungen und Zeitschriften. URL: `http://hdl.handle.net/11022/0000-0000-2E27-2`.

IFLA (2009): *Functional Requirements for Bibliographic Records*. URL: `http://www.ifla.org/publications/functional-requirements-for-bibliographic-records` (besucht am 23. 12. 2016).

Iglesias-Franjo, Estíbaliz und Vilares, Jesús (2016): Searching Four-Millenia-Old Digitized Documents: A Text Retrieval System for Egyptologists. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. 22–31.

IMDI (2003): *Metadata Elements for Session Descriptions: PART 1 A*. URL: `http://www.mpi.nl/ISLE/documents/draft/ISLE_MetaData_2.5.pdf` (besucht am 23. 12. 2016).

IMDI (2009): *ISLE Metadata Initiative (IMDI): PART 1B Metadata Elements for Catalogue Descriptions.* URL: https://tla.mpi.nl/imdi-metadata/attachment/imdi_catalogue_3-0-0/ (besucht am 23. 12. 2016).

Institut für deutsche Sprache Mannheim (2007): *Cosmas II: Digitale Recherche in fünf Millionen Buchseiten.* Hrsg. von Leibniz-Gemeinschaft.

ISO (2014): *Information and documentation – The Dublin Core metadata element set.* URL: http://www.iso.org/iso/catalogue_detail?csnumber=52142 (besucht am 15. 11. 2015).

ISO (2015): *Language resource management. Component Metadata Infrastructure (CMDI).* URL: https://www.iso.org/standard/37336.html (besucht am 26. 06. 2017).

ISO und IEC (2012): *Information technology – Object Management Group Unified Modeling Language (OMG UML) – Part 2: Superstructure.* URL: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=52854 (besucht am 30. 10. 2015).

Jakus, Grega (2013): *Concepts, ontologies, and knowledge representation.* New York: Springer.

Jensen, Uwe; Katsanidou, Alexia und Zenk-Möltgen, Wolfgang (2011): Metadaten und Standards. In: Büttner, Stephan; Hobohm, Hans-Christoph und Müller, Lars (Hrsg.): *Handbuch Forschungsdatenmanagement.* Bad Honnef: Bock + Herchen. 83–100.

Jurafsky, Dan und Martin, James H. (2009): *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.* 2. Aufl. Upper Saddle River: Prentice Hall.

Jurish, Bryan (2010): More than Words: Using Token Context to Improve Canonicalization of Historical German. In: *JLCL* 25(1). 23–40.

Keim, Daniel A. (2016): Die Rolle von Mensch und Computer in den Digital Humanities. In: *DHd 2016. Modellierung Vernetzung Visualisierung.* URL: http://www.dhd2016.de/sounds/Abschluss_DHd_2016.mp4 (besucht am 06. 06. 2016).

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel und Suchomel, Vít (2014): The Sketch Engine: ten years on. In: *Lexicography ASIALEX*(1:7). DOI: 10.1007/s40607-014-0009-9.

Kindling, Maxi und Schirmbacher, Peter (2013): „Die digitale Forschungswelt“ als Gegenstand der Forschung. In: *Information. Wissenschaft & Praxis* 64(2-3). 127–136. DOI: 10.1515/iwp-2013-0017. (Besucht am 13. 01. 2016).

Kirschenbaum, Matthew G. (2004): "So the Colors Cover the Wires": Interface, Aesthetics and Usability. In: Schreibman, Susan; Siemens, Kay und Unsworth, John (Hrsg.): *A Companion to Digital Humanities*. Oxford: Blackwell. 523–542.

Klump, Jens (2009): Digitale Forschungsdaten. In: Neuroth, Heike (Hrsg.): *Nestor-Handbuch*. Boizenburg und Göttingen: Hülsbusch und Univ.-Verl. Göttingen. 104–115.

Koch, Peter und Österreicher, Wulf (1985): Sprache der Nähe – Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. In: *Romanistisches Jahrbuch*(36). 15–43.

Kok, Daniël de; Qiu, Wei und Hinrichs, Marie (2015): WebLicht: Bombardieren bevor die Services explodieren. In: *DHd 2015 Book of Abstracts*.

Kramer, Michael J. (2014): Going Meta on Metadata. In: *Journal of Digital Humanities* 3(2). URL: `http://journalofdigitalhumanities.org/3-2/going-meta-on-metadata/` (besucht am 11. 11. 2015).

Krause, Thomas; Leser, Ulf und Lüdeling, Anke (2016): graphANNIS: A Fast Query Engine for Deeply Annotated Linguistic Corpora. In: *Journal for Language Technology and Computational Linguistics* 31(1). 1–25. URL: `http://www.jlcl.org/2016_Heft1/jlcl-2016-1-1KrauseEtAl.pdf`.

Krause, Thomas; Lüdeling, Anke; Odebrecht, Carolin; Romary, Laurent; Schirmbacher, Peter und Zielke, Dennis (2014): LAUDATIO-Repository: Accessing a heterogeneous field of linguistic corpora with the help of an open access repository. In: *Digital Humanities Conference Abstracts*. 489–490. URL: `http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt` (besucht am 23. 12. 2016).

Krause, Thomas; Lüdeling, Anke; Odebrecht, Carolin und Zeldes, Amir (2012): *Multiple Tokenization in a Diachronic Corpus*. Oslo. URL: `http://www.hf.uio.no/ifikk/english/research/projects/proiel/ealc/` (besucht am 30. 05. 2016).

Krause, Thomas; Lüdeling, Anke; Odebrecht, Carolin und Zielke, Dennis (2015): *LAUDATIO: Ein flexibles Repositorium für historische Textdaten*. Hamburg. URL: `https://www.gwiss.uni-hamburg.de/gwin/ueber-uns/forge2015/programm.html` (besucht am 18. 09. 2016).

Krause, Thomas; Odebrecht, Carolin; Zeldes, Amir und Zipser, Florian (2013): *Unary TEI Elements and the Token Based Corpus*. Rom. URL: `https://www.researchgate.net/publication/261364679_Unary_TEI_Elements_and_the_Token_Based_Corpus`.

Krause, Thomas und Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities* 31(1). 118–139. DOI: 10.1093/llc/fqu057.

Kroch, Anthony und Taylor, Ann (2000): *The Penn-Helsinki Parsed Corpus of Middle English (PPCME2): CD-ROM.* URL: http://www.ling.upenn.edu/ppche-release-2016/PPCME2-RELEASE-4 (besucht am 08. 08. 2016).

Kuebler, Sandra und Zinsmeister, Heike (2015): *Corpus Linguistics and Linguistically Annotated Corpora.* London: Bloomsbury Academic.

Küster, Marc; Ludwig, Christoph und Aschenbrenner, Andreas (2007): TextGrid as a Digital Ecosystem. In: *IEEE DEST.* URL: http://textgrid.de/fileadmin/publikationen/kuester-2007.pdf (besucht am 16. 09. 2016).

Kytö, Merja (1996): *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts: Third Edition.* Helsinki.

Kytö, Merja (2011): Corpora and historical linguistics. In: *Revista Brasileira de Linguistica Aplicada* 11. 417–457. DOI: 10.1590/S1984-63982011000200007.

Leech, Geoffrey (1993): Corpus Annotation Schemes. In: *Literary and Linguistic Computing* 8(4). 275–281.

Lehmann, Nico (2015): *Registervariation zwischen konzeptuell öffentlichem und nicht-öffentlichem Gebrauch gesprochener Sprache: Vortrag im Kolloquium Korpuslinguistik am 18.11.2015.* Berlin. URL: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/lehre/wise-1516/ko_5220053.

Lehmberg, Timm und Wörner, Kai (2008): Annotation standards. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics.* Berlin: De Gruyter. 484–501.

Leipold, Aletta; Kösser, Sylwia; Gießler, André und Solms, Hans-Joachim (2015): Zwischen Online-Korpus und Buch: Die Hybridedition der Wundarznei des Heinrich von Pfalzpaint. In: Bein, Thomas (Hrsg.): *Vom Nutzen der Editionen.* (= Beihefte zu Editio, Bd. 39). De Gruyter. 167–184.

Lemnitzer, Lothar und Zinsmeister, Heike (2006): *Korpuslinguistik: Eine Einführung.* Tübingen: Gunter Narr.

Lickley, Robin J. (2015): Fluency and Disfluency. In: Redford, Melissa A. (Hrsg.): *The Handbook of Speech Production.* Hoboken, NJ: John Wiley & Sons, Inc. 445–474. DOI: 10.1002/9781118584156.ch20.

Lüdeling, Anke (2011): Corpora in Linguistics: Sampling and Annotation. In: Grandin, Karl (Hrsg.): *Going Digital.* (= Nobel Symposium, Bd. 147). New York: Science History Publications. 220–243.

Lüdeling, Anke (2012): A corpus-linguistics perspective on language documentation, data, and the challenge of small corpora. In: Seifart, Frank; Haig, Geoffrey; Himmelmann, Nikolaus P.; Jung, Dagmar; Margetts, Anna und Trilsbeek, Paul (Hrsg.): *Potentials of Language Documentation.* (= Language Documentation & Conservation Special Publication, Bd. 4). Hawaii: University of Hawai'i Press. 32–38.

Lüdeling, Anke; Odebrecht, Carolin und Zeldes, Amir (2014): *RIDGES-Herbology (Version 4.1).* Humboldt-Universität zu Berlin. Register in Diachronic German Science Project. URL: `http://hdl.handle.net/11022/0000-0000-8253-F`.

Lüdeling, Anke; Ritz, Julia; Stede, Manfred und Zeldes, Amir (2016): Corpus Linguistics. In: Fery, Caroline und Ishihara, Shinishiro (Hrsg.): *OUP Handbook of Information Structure.* Oxford: Oxford University Press. 599–617.

Lüdeling, Anke und Zeldes, Amir (2007): Three Views on Corpora: Corpus Linguistics, Literary Computing, and Computational Linguistics. In: *Jahrbuch für Computerphilologie*(9). 149–178.

Lühr, Rosemarie; Faßhauer, Vera; Prutscher, Daniela und Seidel, Henry (2014): *Fuerstinnenkorrespondenz (Version 1.1).* Universität Jena. Frühneuzeitliche Fürstinnenkorrespondenzen im mitteldeutschen Raum. URL: `http://hdl.handle.net/11022/0000-0002-5568-A`.

McCarty, Willard (2004): Modeling: A Study in Words and Meanings. In: Schreibman, Susan; Siemens, Kay und Unsworth, John (Hrsg.): *A Companion to Digital Humanities.* Oxford: Blackwell. 254–270.

McEnery, Tony und Hardie, Andrew (2012): *Corpus Linguistics: Method, Theory and Practice.* (= Cambridge Textbooks in Linguistics). Cambride [u.a.]: Cambridge University Press.

Mengel, Andreas und Lezius, Wolfgang (2000): An XML-based encoding format for syntactically annotated corpora. In: *Proceedings of the Second International Conference on Language Resources and Engineering.* Athen. URL: `http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/TIGERSearch/doc/html/TigerXML.html` (besucht am 01. 12. 2015).

METS Primer (2010): *<METS>: Metadata encoding and transmission standard: primer and reference manual.* URL: `http://www.loc.gov/standards/mets/METSPrimer.doc` (besucht am 16. 09. 2016).

Miller, Steven J. (2011): *Metadata for Digital Collections: A How-To-Do-It Manual.* (= How-To-Do-It Manuals, Bd. 179). New York und London: Neal-Schuman Publishers.

Monachini, Monica; Soria, Claudia und Mapelli, Valerie (2004): *Testing Scenario and Quality Assessment Strategy: D 6.1B*. URL: http://weblic.ilc.cnr.it/viewpage.php/sez=ricerca/id=840/.

Moretti, Franco (2007): *Graphs, Maps, Trees: Abstract Models for Literary History*. London und New York: Verso.

Muhie Yimam, Seid; Gurevych Iryna; Eckart de Castilho, Richard und Biemann, Chris (2013): WebAnno: Flexible, Web-based and Visually Supported System for Distributed Annotations. In: *Proceedings of the 51st Annual Meeting of the Association for Computational*. 1–6.

Neuroth, Heike; Lohmeier, Felix und Smith, Kathleen Marie (2011): TextGrid: Virtual Research Environment for the Humanities. In: *The International Journal of Digital Curation* 6(2). 222–231.

NISO (2004): *Understanding Metadatada*. Hrsg. von National Information Standards Organization. Bethesda. URL: http://www.niso.org/publications/press/UnderstandingMetadata.pdf (besucht am 13. 02. 2015).

NISO (2007): *A Framework of Guidance for Building Good Digital Collections*. Hrsg. von National Information Standards Organization. URL: http://www.niso.org/publications/rp/framework3.pdf (besucht am 05. 08. 2016).

Nivre, Joakim; Hall, John und Nilsson, Jens (2004): Memory-Based Dependency Parsing. In: *Proceedings of the Eighth Conference on Computational Natural Language Learning*. 49–56.

Odebrecht, Carolin (2014): Modeling Linguistic Research Data for a Repository for Historical Corpora. In: *Digital Humanities Conference Abstracts*. 284–285.

Odebrecht, Carolin (2015): Interdisziplinäre Nutzung von Forschungsdaten mithilfe einer technisch-abstrakten Modellierung. In: *DHd 2015 Book of Abstracts*. URL: http://gams.uni-graz.at/o:dhd2015.abstracts-gesamt (besucht am 23. 12. 2016).

Odebrecht, Carolin (2017): MKM - ein Metamodell für Korpusmetadaten: Dokumentation und Wiederverwendung historischer Korpora. Diss. Berlin: Humboldt-Universität zu Berlin.

Odebrecht, Carolin; Belz, Malte; Zeldes, Amir; Lüdeling, Anke und Krause, Thomas (2017): RIDGES Herbology: Designing a Diachronic Multi-Layer Corpus. In: *Language Resources and Evaluation* 51(3). 695–725. DOI: 10.1007/s10579-016-9374-3.

Owens, Trevor (2011): Defining Data for Humanities: Text, Artifacts, Information of Evidence. In: *Journal of Digital Humanities* 1(1).

Perlitz, Laura (2014): Konkurrenz zwischen Wortbildung und Syntax: Historische Entwicklung von Benennung. Diss. Berlin: Humboldt-Universität zu Berlin. URL: http://edoc.hu-berlin.de/master/perlitz-laura-2014-08-08/PDF/perlitz.pdf (besucht am 16. 06. 2016).

Petrova, Svetlana; Solf, Michael; Ritz, Julia; Chiarcos, Christian und Zeldes, Amir (2009): Building and using a Richly Annotated Interlinear Diachronic Corpus: The Case of Old High German Tatian. In: *Traitement Automatique des Langues, ATALA* 50(29). 47–71.

Piotrowski, Michael (2012): *Natural Language Processing for Historical Texts.* (= Synthesis Lectures on Human Language Technologies, Bd. 17). San Rafael: Morgan & Claypool.

Pitti, Daniel (2004): Designing Sustainable Projects and Publications. In: Schreibman, Susan; Siemens, Kay und Unsworth, John (Hrsg.): *A Companion to Digital Humanities.* Oxford: Blackwell. 471–487. URL: http://digitalhumanities.org/companion/view?docId=blackwell/9781405103213/9781405103213.xml%5C&chunk.id=ss1-5-1 (besucht am 09. 09. 2016).

Qin, Jian und Li, Kai (2013): How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. In: DCMI (Hrsg.): *DC-2013–The Lisbon Proceedings.* 25–34. URL: http://dcpapers.dublincore.org/pubs/issue/view/165 (besucht am 23. 12. 2016).

R Core Team (2016): *R: A language and environment for statistical computing.* Wien.

Reznicek, Marc; Lüdeling, Anke; Krummes, Cédric; Schwantuschke, Franziska; Walter, Maik; Schmidt, Karin; Hirschmann, Hagen und Torsten, Andreas (2012): *Falko-Handbuch: Korpusaufbau und Annotation Version 2.01. Technical Report: Humboldt-Universität zu Berlin.* Berlin.

Riley, Jenn und Becker, Devin (2009): *Seeing Standards: A Visualization of the Metadata Universe.* URL: http://jennriley.com/metadatamap/ (besucht am 07. 01. 2017).

Romary, Laurent (2012): *Topics in data modeling, linguistic annotation and digital libraries: KAIST Global Lectures, 17-30 January 2012.* Daejeon.

Romary, Laurent (2013): *Technologies, services and user expectations: Prospects for DARIAH. R EU 4.3.2. Research Report. State and University Library Goettingen.* URL: https://hal.inria.fr/hal-00912653 (besucht am 10. 01. 2017).

Romary, Laurent und Chambers, Sally (2014): *DARIAH: Advancing a digital revolution in the arts and humanities across Europe: Data Archiving and Networked Services (DANS).*

Romary, Laurent und Ide, Nancy (2004): International standard for a linguistic annotation framework. In: *Natural Language Engineering* 10(3-4). 211–225. DOI: 10.1017/S135132490400350X.

Romary, Laurent und Tucnak, Zina (2002): Experience with OLAC for the ATILF archives. In: *Third International Conference on Language Resources and Evaluation.* URL: https://hal.inria.fr/inria-00101041.

Romary, Laurent und Witt, Andreas (2012): Data Formats for Phonological Corpora. In: Durand, Jacques; Gut, Ulrike und Kristoffersen, Gjert (Hrsg.): *Handbook of corpus phonology.* Oxford: Oxford University Press. 211–225.

Romary, Laurent; Zeldes, Amir und Zipser, Florian (2015): <tiger2/>: serialising the ISO SynAF syntactic object model. In: *Language Resources and Evaluation* 49(1). 1–18. DOI: 10.1007/s10579-014-9288-x.

Rümpel, Stefanie (2011): Der Lebenszyklus von Forschungsdaten. In: Büttner, Stephan; Hobohm, Hans-Christoph und Müller, Lars (Hrsg.): *Handbuch Forschungsdatenmanagement.* Bad Honnef: Bock + Herchen. 25–34.

Rupp, Chris; Hahn, Jürgen; Queins, Stefan; Jeckle, Mario und Zengler, Barbara (2005): *UML 2 glasklar: Praxiswissen für die UML-Modellierung und -Zertifizierung.* 2. Auflage. München und Wien: Hanser.

Sahle, Patrick und Kronenwett, Simone (2013): Jenseits der Daten: Überlegungen zu Datenzentren für die Geisteswissenschaften am Beispiel des Kölner 'Data Center For The Humanities'. In: *Library Ideas*(23). 76–96.

Salmon-Alt, Susanne; Romary, Laurent und Pierrel, Jean-Marie (2006): Un modèle générique d'organisation de corpus en ligne: application à la FReeBank. In: *Traitement Automatique des Langues, ATALA*(45). 145–169.

Sauer, Simon und Lüdeling, Anke (2016): Flexible Multi-Layer Spoken Dialogue Corpora. In: *International Journal of Corpus Linguistics. Special Issue on Spoken Corpora.* 419–438. DOI: 10.1075/ijcl.21.3.06sau.

Schäfer, Roland und Bildhauer, Felix (2012): Building Large Corpora from the Web Using a New Efficient Tool Chain. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* European Language Resources Association (ELRA). 486–493.

Schiller, Anne; Teufel, Simone; Stöckert, Christine und Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textkorpora mit STTS.* Hrsg. von Universität Tübingen. Seminar für Sprachwissenschaft. Tübingen.

Schmid, Helmut (2008): Tokenization and Part-of-speech Tagging. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics.* Berlin: De Gruyter. 527–551.

Schmid, Helmut und Laws, Florian (2008): Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In: *Proceedings of the 22nd International Conference on Computational Linguistics (COLING).* 777–784.

Schmidt, Thomas (2011): A TEI-based approach to standardising spoken language transcription. In: *Journal of the Text Encoding Initiative*(1). DOI: 10.4000/jtei.142.

Schmidt, Thomas und Wörner, Kai (2009): EXMARaLDA: Creating, analysing and sharing spoken language corpora for pragmatic research. In: *Pragmatics* 19(4). 565–582.

Schroeder, Carolin T.; Zeldes, Amir und et al. (2016): *Coptic SCRIPTORIUM, 2013-2016.* URL: http://copticscriptorium.org (besucht am 16. 03. 2016).

Schweitzer, Antje und Lewandowski, Natalie (2010): *Gesprächscorpus.* Universität Stuttgart. SFB 732 Projekt A4. URL: http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/IMS-GECO.html.

Schweitzer, Antje und Lewandowski, Natalie (2013): Convergence of Articulation Rate in Spontaneous Speech. In: *Proceedings of Interspeech.* 525–529.

Simanowski, Roberto (2011): *Digital Art and Meaning: Reading Kinetic Poetry, Text Machines, Mapping Art, and Interactive Installations (Electronic Mediations).* Minnesota: University of Minnesota Press.

Simons, Gary (2014): The role of metadata in the infrastructure for archival interoperation. In: *Language and Linguistics Compass* 8(11). 486–494. DOI: 10.1111/lnc3.12126.

Simons, Gary und Bird, Steven (2008): Toward a global infrastructure for the sustainability of language resources. In: *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation.* 87–100.

Sinclair, John (1995): *Corpus, concordance, collocation.* 3. Auflage. (= Describing English language). Oxford: Oxford University Press.

Solms, Hans-Joachim und Wegera, Klaus-Peter (1998): Das Bonner Frühneuhochdeutsch-Korpus: Rückblick und Perspektiven. In: Bergmann, Rolf (Hrsg.): *Probleme der Textauswahl für einen elektronischen Thesaurus. Beiträge zum ersten Göttinger*

Arbeitsgespräch zur historischen deutschen Wortforschung, 1. und 2. November 1996, Stuttgart. Leipzig. 22–39. URL: https://korpora.zim.uni-duisburg-essen.de/fnhd/ (besucht am 03. 08. 2016).

Solodovnik, Iryna (2011): Metadata issues in Digital Libraries: key concepts and perspectives. In: *Italian Journal of Library, Archives and Information Science* 2(2). 1–27. DOI: 10.4403/jlis.it-4663.

Stiller, Juliane; Thoden, Klaus und Zielke, Dennis (2016): Usability in den Digital Humanities am Beispiel des LAUDATIO-Repositoriums. In: *DHd 2016. Modellierung Vernetzung Visualisierung.* 244–247.

Stührenberg, Maik (2012): The TEI and Current Standards for Structuring Linguistic Data. In: *Journal of the Text Encoding Initiative*(3). 1–14.

TEI Consortium (2015): *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* URL: http://www.tei-c.org/Guidelines/P5/ (besucht am 15. 11. 2015).

Telljohann, Heike; Hinrichs, Erhard W. und Kübler, Sandra (2003): *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z): Technical report, Seminar für Sprachwissenschaft.* URL: http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf (besucht am 06. 01. 2016).

Van Uytvanck, Dieter; Goosen, Twan und Windhouwer, Menzo (2012): *CMDI and granularity.* URL: https://www.clarin.eu/sites/default/files/AP3-007-CMDI_and_granularity.pdf (besucht am 07. 01. 2017).

Van Uytvanck, Dieter; Stehouwer, Herman und Lempen, Lari (2012): Semantic metadata mapping in practice: the Virtual Language Observatory. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* European Language Resources Association (ELRA). 1029–1033.

van Zundert, Joris (2012): If you build it, will we come? Large scale digital infrastructures as a dead end for digital humanities. In: *Historical Social Research – Historische Sozialforschung* 37(3). 165–186.

Vertan, Cristina; Ellwardt, Andreas und Hummerl, Susanne (2016): Ein Mehrebenen-Tagging-Modell für die Annotation altäthiopischer Texte. In: *DHd 2016. Modellierung Vernetzung Visualisierung.* 258–261.

Voigt, Vivian; Zipser, Florian und Odebrecht, Carolin (2016): *SaltInfoModule: Automatische Metadatenextraktion und Dokumentation für Korpora: DGfS-CL Poster Session. 38. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft (DGfS). Poster.* Konstanz.

Weddige, Hilkert (2006): *Einführung in die germanistische Mediävistik.* München: Beck.

Wegera, Klaus-Peter (2013): Language data exploitation: design and analysis of historical language corpora. In: Bennett, Paul; Durrell, Martin; Scheible, Silke und Whitt, Richard J. (Hrsg.): *New Methods in Historical Corpora.* (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache, Bd. 3). Tübingen: Narr. 55–73.

Wichmann, Anne (2008): Spoken corpora and speech corpora. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics.* Berlin: De Gruyter. 187–207.

Wiese, Lena (2015): *Advanced Data Management: For SQL, NOSQL, Cloud and Distributed Databases.* Oldenburg: De Gruyter.

Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, I Jsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientific data* 3. 160018. DOI: 10.1038/sdata.2016.18.

Windhouwer, Menzo (2012): RELcat: a Relation Registry for ISOcat data categories. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* European Language Resources Association (ELRA). 3661–3664.

Wittenburg, Peter; Brugmann, Hennie; Russel, Albert; Klassmann, Alex und Sloetjes, Han (2006): ELAN: A Professional Framework for Multimodality Research. In: *Proceeedings of LREC.* 1556–1559. URL: http://www.lrec-conf.org/proceedings/lrec2006/ (besucht am 23. 12. 2016).

Wittenburg, Peter; Gibbon, Dafydd und Peters, Wim (2001): *Metadata Elements for Lexicon Descriptions: Technical Report PART 1C.* URL: www.mpi.nl/ISLE/documents/draft/ISLE_Lexicon_1.0.pdf (besucht am 13. 01. 2017).

Wittenburg, Peter; Mosel, Ulrike und Dwyer, Arienne (2002): Methods of language documentation in the DOBES project. In: Wittenburg, Peter (Hrsg.): *Proceedings of the International Workshop on Resources and Tools in Field Linguistics.* URL: www.mpi.nl/lrec/2002/papers/lrec-pap-02b-dobes-talk-final.pdf (besucht am 13. 01. 2017).

Xie, Iris und Matusiak, Krystyna (2016): *Discover Digital Libraries: Theory and Practice.* Oxford: Elsevier. DOI: 10.1016/B978-0-12-417112-1.00005-3.

Zeldes, Amir (erscheint): The Case for Caseless Prepositional Constructions with "voller" in German. In: Boas, Hans C. und Ziem, Alexander (Hrsg.): *Constructional Approaches to Syntactic Structures in German.* (= Trends in Linguistics: Studies and Monographs.). Berlin: De Gruyter.

Zeldes, Amir und Schroeder, Carolin T. (2015): Computational Methods for Coptic: Developing and Using Part-of-Speech Tagging for Digital Scholarship in the Humanities. In: *Digital Scholarship in the Humanities* 31(1). 164–176.

Zeng, Marcia Lei und Qin, Jian (2016): *Metadata.* Second Edition. London: facet publishing.

Zinsmeister, Heike; Hinrichs, Erhard W.; Kübler, Sandra und Witt, Andreas (2008): Linguistically annotated corpora: Quality assurance, reusability and sustainability. In: Lüdeling, Anke und Kytö, Merja (Hrsg.): *Corpus Linguistics.* Berlin: De Gruyter. 759–776.

Zipser, Florian (2009): *Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells.* Berlin. URL: `hal-00606102` (besucht am 25. 09. 2016).

Zipser, Florian (2014): *SaltNPepper und das Formatpluriversium.* Berlin. URL: `http://www.laudatio-repository.org/laudatio/wp-admin/tmp/2014/11/saltNpepper_laudatio_2014_FZ.pdf`.

Zipser, Florian und Romary, Laurent (2010): A model oriented approach to the mapping of annotation formats using standards. In: *Proceedings of the Workshop on Language Resource and Language Technology Standards, LREC 2010.* URL: `http://hal.archives-ouvertes.fr/inria-00527799/en/` (besucht am 12. 11. 2014).